Attachment A

Application Cover Sheet

<table>
<tr><td colspan="4">Name of Respondent – CLOUDSUFI, Inc.</td></tr>
<tr><td>Mailing Address –
333 West San Carlos Street
Riverpark Tower, Ste 600</td><td>City/Town –
San Jose</td><td>State –
CA</td><td>Zip Code –
95110</td></tr>
<tr><td>Telephone – (408) 605-4035</td><td colspan="2">Fax</td><td>Web Address – cloudsufi.com</td></tr>
<tr><td colspan="2">Primary Contact for Clarification – Irfan Khan</td><td colspan="2">Primary Contact E-mail Address –
irfan@cloudsufi.com</td></tr>
<tr><td colspan="2">Authorized Signatory – Irfan Khan</td><td colspan="2">Authorized Signatory E-mail Address –
irfan@cloudsufi.com</td></tr>
<tr><td colspan="2">Legal Status/Jurisdiction – Delaware Corporation</td><td colspan="2">Respondents W-9
Respondents UEI No. –
84-1919490</td></tr>
</table>

Attachment B

Massachusetts Technology Collaborative
Authorized Respondent's Signature and Acceptance Form

The undersigned is a duly authorized representative of the Respondent listed below. The Respondent has read and understands the RFP requirements. The Respondent acknowledges that all of the terms and conditions of the RFP are mandatory, and that Respondent's response is compliant with such requirements.

The Respondent understands that, if selected by the Mass Tech Collaborative, the Respondent and the Mass Tech Collaborative will execute an Agreement specifying the mutual requirements of participation.

Respondent agrees that the entire bid response will remain valid for sixty (60) days from receipt by the Mass Tech Collaborative.

I certify that Respondent is in compliance with all corporate filing requirements and State tax laws.

I certify that the statements made in this response to the RFP, including all attachments and exhibits, are true and correct to the best of my knowledge.

I acknowledge and understand the procedures for handling confidential materials submitted to the Mass Tech Collaborative as set forth in Section 3.1(d); I agree to be bound by those procedures and agree that the Mass Tech Collaborative shall not be liable under any circumstances for the disclosure of any materials submitted to the Mass Tech Collaborative pursuant to this RFP or upon the Respondent's selection.

Respondent: **CLOUDSUFI, Inc.**
(Printed Name of Respondent)

By:
(Signature of Authorized Representative)

Name: **Irfan Khan**
Title: **President & CEO**
Date: **September 15, 2025**

Proposals for

AI Hub Data Commons Collaborative

CONSULTING SERVICES

RFP No. 2026-JAII-01

Massachusetts Technology Collaborative

75 North Drive
Westborough, MA 01581-3340
http://www.masstech.org

Submitted by : CLOUDSUFI Inc.

07 OCT 2025

Table of Contents

I. Executive Summary & Partnership Vision **2**
IA. Why Google cloud ?... Google's Work in Data Commons 4
1. Data Integration and Standardization 4
2. Tools and Accessibility 4
3. AI Grounded in Facts 5
4. Leader in Artificial Intelligence 5
5. Commitment from the Extended Google Team 5
6. Google Executive Briefings 6
II. Profile & Qualification **7**
1. Expertise in: Data Commons / Cataloging Platforms (CKAN, Dataverse) 9
2. Expertise in: Metadata Management and FAIR Standards 9
3. Expertise in: Synthetic Data (tools like SDV, Synthea, CARLA) 10
4. Expertise in: AI Fairness (AIF360, Fairlearn) 10
5. Expertise in: Security/Governance Frameworks (RBAC, consent management, NIST standards) 11
6. Google Expertise: Unique Characteristics of Google 12
More than just world-class infrastructure 12
III. Proposed Solution Architecture: The Integrated Data Commons **15**
IV. Technical Capabilities: Core Infrastructure & Cataloging **19**
1. Presentation and Access Layer 20
2. Control Plane & Governance 21
3. Data Core & Curation Engine 22
4. Central Storage & MLOps Platform (Zone A) 22
5. Zone B: Hosted/On-Prem - High-Performance Compute & AI 23
V. Enabling Financial Governance: The Commons Credits System **26**
VI. Differentiating AI Capabilities **29**
VI.A. Synthetic Data Generation (RFP 2.4) 29
VI.B. Responsible AI and Bias Framework (RFP 2.5) 30
VI.C Deeply Embedded in Google Cloud's Innovation Efforts 31
VII. Security, Governance, and Compliance Framework **32**
VIII. Project Management, Delivery, and Ecosystem Growth **34**
Detailed Project Plan (40 Weeks) 36
Project Governance & Change Management 43
RFP Requirements Compliance Checklist (Phase Integration) 44
IX. RFP Requirements Compliance Checklist **46**
X. Assumptions **49**
XI. Past Performance & References **52**
Knowledge Fabric for Technology Disruption Framework 52
Patient 360 Platform 53
Industry : Fashion and Jewelry 54
XII. Bios of Key Personnel **56**

I. Executive Summary & Partnership Vision

CLOUDSUFI, leveraging the power and proven architecture of **Google's Data Commons** and partnering with **Google Cloud Platform (GCP)**, proposes a strategic partnership to establish a world-class **Data Commons Collaborative (DCC)** for the Massachusetts AI Hub. This solution is designed for **rapid implementation, superior security, accessibility, and unprecedented value**, directly addressing the RFP's core evaluation criteria.

The foundation of our proposal is **Google Cloud's operational Data Commons infrastructure**—the only cloud-native knowledge graph of its kind among major providers. Rather than building a bespoke platform from scratch, we will deliver a **secure, customized instance** of this already proven technology. This approach accelerates the timeline, aligns with Massachusetts' vision for a statewide Data Commons, and reduces the risks that typically arise in large-scale data integration and harmonization efforts.

To ensure precision, transparency, and optimal value for the Commonwealth, **CLOUDSUFI and Google Cloud are jointly proposing an initial investment phase - two weeks of on-site workshops at zero cost to MassTech.** During this phase, our combined teams will work closely with stakeholders to document existing systems, validate data flows, and refine infrastructure sizing and cost models based on real-world conditions. By absorbing this upfront effort, we **de-risk implementation** for the Commonwealth and ensure that subsequent cost estimates reflect verified requirements rather than assumptions—demonstrating our **commitment to partnership, accountability, and long-term success.**

The proposed platform extends far beyond data aggregation. We deliver a **cohesive, serverless, and secure ecosystem** built on best-in-class, natively integrated Google services:

- **BigQuery** for petabyte-scale analytics
- **Vertex AI with Gemini models** for advanced modeling and synthetic data
- **Apigee** for secure API management with **API keys, rate-limits, and quotas** to ensure fair access
- **Dataplex** for governance and metadata
- **Looker** for transparency, reporting, and dashboards

This integrated stack provides lower total cost of ownership, higher scalability, and less operational risk than fragmented, open-source alternatives (e.g., CKAN, Dataverse).

The CLOUDSUFI–Google partnership also brings **unique credibility**. We are actively co-developing **governance capabilities on Dataplex** and hold the **three-year Google Data Commons Data Loading SOW**, underscoring our operational role in the same knowledge graph Massachusetts will adopt. This

ensures that the Commonwealth is not betting on untested tools but aligning with the **roadmap powering Google's global data innovation**.

Critically, the DCC will transform into a **living knowledge graph** with **semantic search and natural language (NLP) interfaces available in Phase 1**, allowing researchers, policymakers, and innovators to pose questions in everyday language rather than navigate schemas. Coupled with Google's **Responsible AI (RAI) toolkit** for fairness, explainability, and bias mitigation, and our **Commons Credits system** for transparent, equitable allocation of resources, the platform ensures broad participation and long-term sustainability.

Together, **Google Cloud and CloudSufi** will deliver not just a system that meets the RFP's requirements, but a **strategic asset that positions Massachusetts as a national leader in responsible, AI-driven innovation and economic development.**

Compliance Highlights at Launch: The DCC will provide **semantic search and natural-language access in Phase 1**, operate across **Dev/Test, QA, and Production** environments, and expose a secure **API gateway (Apigee) with API keys, rate-limits, quotas, and monitoring**. The public web experience will meet **WCAG 2.1 Level AA** with support for screen readers, keyboard navigation, and captioning. Audit logs will be retained for **at least three years** with options for archival expansion as policies are finalized.

IA. Why Google cloud ?... Google's Work in Data Commons

Google created and maintains Data Commons as an open-source platform whose mission is to organize the world's publicly available statistical data and make it universally accessible and useful.
Google's work for Data Commons is to act as a central data integrator, transformer, and distribution point, making complex, siloed public information easy to access and leverage, especially with AI.
Outlined below are three main areas summarizing the work that Google does for Data Commons.

1. Data Integration and Standardization

The most fundamental work Google does is taking vast amounts of fragmented public data and uniting it into a cohesive resource. A few key activities include:

- Creates a Unified Knowledge Graph: We collect data from over 200 trusted sources, including the U.S. Census Bureau, the World Bank, the United Nations, and various government and non-governmental organizations (NGOs).
- Data Normalizing and Cleaning: We perform the tedious, costly work of normalizing, cleaning, and reconciling disparate data. This means aligning different schemas, formats, and references so that data about the same entity from different sources can be accessed instantly through a single query.
- Non-Ownership: Lastly, we do not collect or own any of the data itself - we simply aggregate, process, and present publicly available information and provide the link back to the original source.

2. Tools and Accessibility

We develop multiple user-friendly tools to make the data accessible to everyone. We do this by:

- APIs and Libraries: We provide programmatic access via REST APIs and a Python client library and this enables data scientists and developers to integrate Data Commons into their custom projects.
- Web-Based Explorers: For non-coders this provides tools on the website (e.g. Map Explorer, Statistical Variable Explorer, and Scatter Plot Explorer) to analyze, visualize, and compare data without writing code.
- Google Search Integration: A subset of Data Commons' data is directly integrated into Google Search. When a user asks a statistical question (e.g., "What is the population of New Zealand?"), the factual answer, often presented as a knowledge graph or chart, is retrieved directly from Data Commons.

3. AI Grounded in Facts

In recent years, we have focused on using Data Commons as a trusted "data backbone" for our AI models.

- Mitigating AI Inconsistencies: Data Commons is used to ground Large Language Models (LLMs) in real-world, verifiable statistical data. This technique helps reduce the rate of "hallucinations" (inaccurate or invented information) by providing the model with sourced facts.
- Natural Language Querying: We developed the Model Context Protocol (MCP) Server to allow AI agents to consume Data Commons natively. This means developers can query the vast dataset using simple, natural language questions, without having to master complex APIs.
- AI Model Training: Models like our open DataGemma are specifically built to leverage the Data Commons knowledge graph to enhance their factuality and reasoning. This serves as an explicit bridge between LLMs and real-world data.

4. Leader in Artificial Intelligence

Google holds a prominent leadership position in Artificial Intelligence (AI), driven by decades of foundational research and the creation of cutting-edge AI models. Our leadership is anchored by Google DeepMind, which is responsible for landmark breakthroughs like AlphaGo and the protein-folding system AlphaFold, alongside the development of its most powerful and multimodal large language model, Gemini. Google has integrated our AI at every level, from its specialized Tensor Processing Units (TPUs) for accelerated model training, to core product features like AI Overviews in Search, and the Vertex AI platform, that allows enterprises to build and deploy their own AI solutions.

5. Commitment from the Extended Google Team

We are committed to bring ‘the best of Google’ to support the Massachusetts Technology Collaborative and the Data Commons project. Outlined below in Figure 1 are some of the key resources that Google will bring to the partnership.

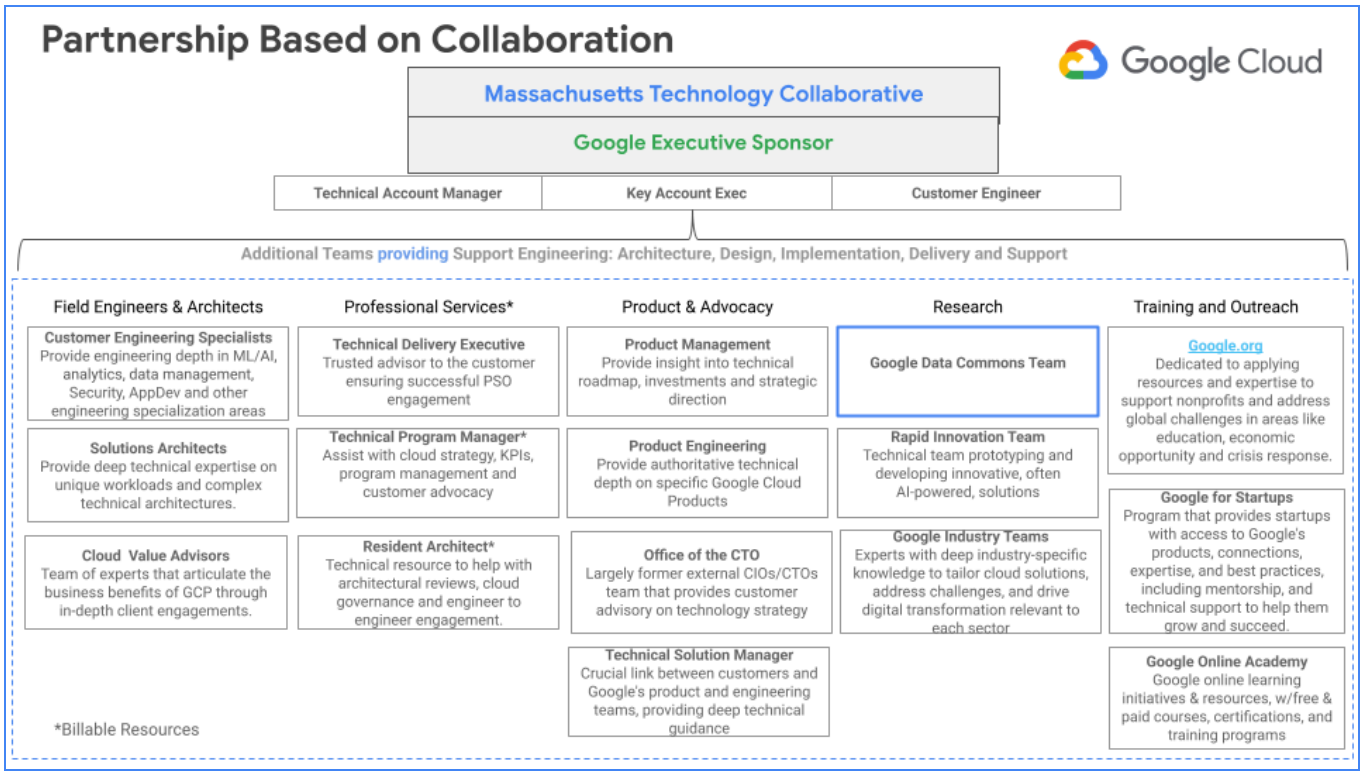

Partnership Based on Collaboration
Google Cloud
Massachusetts Technology Collaborative
Google Executive Sponsor
Technical Account Manager
Key Account Exec
Customer Engineer
Additional Teams providing Support Engineering: Architecture, Design, Implementation, Delivery and Support
Field Engineers & Architects
Professional Services*
Product & Advocacy
Research
Training and Outreach
Customer Engineering Specialists
Provide engineering depth in ML/AI, analytics, data management, Security, AppDev and other engineering specialization areas
Technical Delivery Executive
Trusted advisor to the customer ensuring successful PSO engagement
Product Management
Provide insight into technical roadmap, investments and strategic direction
Google Data Commons Team
Google.org
Dedicated to applying resources and expertise to support nonprofits and address global challenges in areas like education, economic opportunity and crisis response.
Solutions Architects
Provide deep technical expertise on unique workloads and complex technical architectures.
Technical Program Manager*
Assist with cloud strategy, KPIs, program management and customer advocacy
Product Engineering
Provide authoritative technical depth on specific Google Cloud Products
Rapid Innovation Team
Technical team prototyping and developing innovative, often AI-powered, solutions
Google for Startups
Program that provides startups with access to Google's products, connections, expertise, and best practices, including mentorship, and technical support to help them grow and succeed.
Cloud Value Advisors
Team of experts that articulate the business benefits of GCP through in-depth client engagements.
Resident Architect*
Technical resource to help with architectural reviews, cloud governance and engineer to engineer engagement.
Office of the CTO
Largely former external CIOs/CTOs team that provides customer advisory on technology strategy
Google Industry Teams
Experts with deep industry-specific knowledge to tailor cloud solutions, address challenges, and drive digital transformation relevant to each sector
Technical Solution Manager
Crucial link between customers and Google's product and engineering teams, providing deep technical guidance
Google Online Academy
Google online learning initiatives & resources, w/free & paid courses, certifications, and training programs
*Billable Resources

Figure 1. Google Team Supporting the Data Commons Project in addition to CLOUDSUFI Inc.

6. Google Executive Briefings

As part of the partnership, Google welcomes the opportunity to host an annual executive briefing with Massachusetts Technology Collaborative staff, CLOUDSUFI, and the extended Google team. These briefings can serve as a platform to bring us together to discuss technology developments that can be leveraged to enhance the data commons project, product roadmap enhancements, and other key topics that are important to the parties.

II. Profile & Qualification

CLOUDSUFI Profile	
SECTION	DETAILS
Company Name	Cloudsufi Inc.
Headquarters	San Jose, California, United States
Founded	2019
Legal Status	LLC
Website	www.cloudsufi.com
Vision	Eliminate the gap between “Human Intuition” and “Data-driven Decisions
Core Competencies	- Product Engineering - Platform, Mobile , Web - AI , ML & Data Engineering - Advanced Analytics - Data Commons & Metadata Catalogs - AI Fairness & Responsible AI Frameworks - Cloud-Native Architecture & Governance - Security & Compliance (RBAC, MFA, HIPAA, SOC2, FedRAMP experience)
Strategic Partnerships	- Google Cloud (Preferred Technology Partner) - Workato, Oracle,SAP
Certifications	- Google Cloud Premier Partner - Google Cloud Certified Engineers & Architects - ISO/SOC2 readiness and security compliance experience
Relevant Experience	- Delivered large-scale data platforms - Implemented AI-driven synthetic data solutions - Built governance frameworks with FAIR-compliant data access

CLOUDSUFI Profile	
SECTION	DETAILS
Leadership & Team	- Program Managers with experience in multi-stakeholder AI/Cloud programs - Google Cloud–certified Architects & Engineers - Specialists in governance, fairness, and compliance
Differentiators	- Exclusive alignment with Google Cloud AI & Data services - Strong expertise in ethical AI and bias mitigation - Proven phased delivery methodology - Cost-efficient, scalable cloud-native designs
Corporate Responsibility	- Commitment to responsible AI and inclusion - Supports sustainable cloud operations - Focus on equitable access to AI innovation in Massachusetts and beyond

Name of Technology Partner & Sponsor – **Google Cloud**	
Customer	Massachusetts Technology Collaborative
RFP Title	AI Hub Data Commons Collaborative Consulting Services
RFQ No.	2026-JAII-01
Submitted by	Annette Ball
Phone No.	(518) 429-4896
Email	annetteball@google.com
Capture Manager	Mark Slafka | mslafka@google.com | (703) 483-1416

CLOUDSUFI with Google Cloud approaches the DCC not merely as a technology implementation but as the creation of a sophisticated, extensible platform designed to transform data silos into a knowledge-sharing ecosystem. Our proposal focuses on leveraging best-in-class technologies, whether open-source or proprietary, to deliver a solution that meets the Commonwealth's goals for scalability, security, and ethical AI development.

1. Expertise in: Data Commons / Cataloging Platforms (CKAN, Dataverse)

The RFP requires the implementation of **metadata management, cataloging, and discovery tools**, explicitly suggesting that the **CKAN** and **Dataverse** platforms be considered.

- **RFP Context:** CLOUDSUFI understands that the tools listed (including CKAN, Dataverse, DataHub, and Amundsen) are provided as **illustrative references** and are **not mandated solutions**. Vendors are encouraged to propose platforms that are the best fit based on criteria like scalability and interoperability.
- **CLOUDSUFI Approach:** Our strategy focuses on designing a superior, deeply integrated metadata solution that provides the required capability for dataset discovery. We will propose an approach that supports both **auto-harvesting from external sources** (e.g., existing CKAN/Dataverse catalogs) and rigorous **manual curation** workflows to ensure quality and governance. The resulting architecture will support a dataset directory that is **searchable and filterable** by key criteria such as data type, domain, source, and file format.

2. Expertise in: Metadata Management and FAIR Standards

A core requirement of the DCC is ensuring **metadata completeness** and rigorous adherence to **FAIR (Findable, Accessible, Interoperable, Reusable) data principles**.

- **FAIR Implementation:** CLOUDSUFI recognizes that compliance with FAIR principles, supported by widely used frameworks such as **DCAT** and **Schema.org**, is essential to ensure that datasets are discoverable and reusable.
- **Metadata Scope:** Our solution will manage the three required categories of metadata: **Descriptive Metadata** (Title, license), **Technical Metadata** (File format, schema), and **Provenance** (who processed the data).
- **Lineage and Versioning:** We will ensure that the initial launch includes, at a minimum, **dataset-level lineage tracking** to capture the origin and transformations of the entire dataset. Our roadmap will include the capability to incorporate more granular **column-level lineage** in future phases to meet evolving transparency needs. Metadata records themselves will also be **versioned** to maintain a history of schema and description changes.

3. Expertise in: Synthetic Data (tools like SDV, Synthea, CARLA)

The RFP mandates the integration of a **Synthetic Data Capability** to support model training and testing in privacy-preserving and scalable ways.

- **Tooling:** CLOUDSUFI is highly experienced with the open-source tools cited—**SDV** (for tabular data), **Synthea** (for health records), and **CARLA** (for simulations)—but recognizes these are provided only as **illustrative examples**. We are prepared to recommend the most appropriate tools and frameworks (custom, existing, or generative AI-based) to meet the DCC's needs.
- **Scope and Priority:** The synthetic generation capability must support **multimodal** data (image/video/audio) and allow for **configurable privacy-utility tradeoffs**. Priority domains for synthetic data pilots include **life sciences, robotics, education, and healthcare**.
- **Governance:** The platform must include **audit trails** linking synthetic datasets back to the original real datasets for validation and compliance purposes. Furthermore, the pipelines must align with applicable standards such as **HIPAA, GDPR, and relevant Massachusetts state privacy laws**. MassTech is open to full-scale synthetic derivative datasets for production AI applications, not just demonstration sets.

4. Expertise in: AI Fairness (AIF360, Fairlearn)

Ensuring ethical and equitable AI development requires a robust fairness and bias framework, which must include implementation of **bias detection, auditing, and mitigation tools**.

- **Tooling:** Like the synthetic data tools, **AIF360** and **Fairlearn** are listed as illustrative examples of the required capability. CLOUDSUFI's expertise encompasses not just these specific libraries but a comprehensive toolkit that supports **dataset auditing, bias mitigation, model explainability, and drift detection**.
- **Scope:** The platform must support fairness checks for both **models and datasets**. The Massachusetts AI Hub is concerned with a broad range of biases (data, algorithmic, societal) and requires evaluation across protected attributes such as **race, gender, age, and disability**.
- **Monitoring and Reporting:** The solution must support the **continuous monitoring of fairness drift in deployed models**. For this phase of work, fairness reporting is intended primarily for **internal auditing and compliance**, though public transparency features will be supported. Governance should align with standards such as **NIST AI Risk Management Framework**.

5. Expertise in: Security/Governance Frameworks (RBAC, consent management, NIST standards)

Security is the foundation of the DCC, demanding adherence to state and federal requirements, including robust access control and auditability.

- **Access Control (RBAC & MFA):** We will design and implement the required **Role-Based Access Control (RBAC)** to manage access across the required dataset tiers (Public, Restricted, Internal). All sensitive data access will be protected by **Multi-Factor Authentication (MFA)**.
- **Data Residency and Security:** The architecture will ensure that all data is **encrypted at rest and in transit**. All data processing must occur in **U.S. regions with an East-coast preference**, and remote access to Commonwealth Data from outside the continental U.S. is strictly **prohibited**. The DCC must adhere to **MA CISO data classification security standards**.
- **Consent Management:** CLOUDSUFI will implement the required dynamic and verifiable **consent management system in Phase 1**. This system will ensure that consent, primarily for restricted datasets, is **tokenized, revocable, and time-bound**.
- **Auditing and Compliance:** The system is required to **log all data access, downloads, and synthetic generation events**. These audit logs must be retained for at least **three years** for compliance review. We will propose security architectures that align with widely recognized best practices and can adapt to frameworks like the **NIST AI Risk Management Framework**.

6. Google Expertise: Unique Characteristics of Google

More than just world-class infrastructure

- **Infrastructure as Scale**

  The Massachusetts Technology Collaborative will leverage Google's unparalleled backbone network. This network, which moves at "Google speed" — near-zero latency — for billions of users worldwide is now available to governments and enterprises.

- **Advanced Analytics**

  We bring more than just infrastructure, but distinct capabilities in leveraging unparalleled leadership in advanced analytics and intelligent automation, built upon foundational technologies that redefine how state agencies operate.

  We offer world-class capability to unify and extract deep, actionable insights from vast and diverse government datasets, transforming raw information into actionable intelligence that fuels smarter policy and more effective program delivery. Furthermore, our pioneering approach to highly scalable and resilient application environments provides a robust foundation for developing agile and future-proof digital services, ensuring continuous delivery even as demands evolve.

- **Google Cloud Security**

  As a cloud innovator, Google understands security in the cloud. Our cloud services are designed to deliver better security than many on-premises approaches. We make security a priority in our operations—operations that serve users across the world. Security drives our organizational structure, culture, training priorities, and hiring processes. It shapes the design of our data centers and the technology that they house. It's central to our everyday operations and disaster planning, including how we address threats. It's prioritized in the way we handle customer data, our account controls, our compliance audits, and our certifications.

- **Commitment to Sustainability**

  Google Cloud is committed to leading the technology industry in sustainability, not as a separate initiative, but as a core tenet of our business. Our strategy is built on a foundation of ambitious goals, operational excellence, and a commitment to empowering our customers to achieve their own sustainability objectives. We believe that a sustainable cloud is a better cloud—more efficient, more resilient, and more responsible.

- **Proud to call Cambridge Home**

  Our Cambridge office has employees working across more than 30 teams, including our Global Business Organization (sales) and many different engineering and product areas. Some of the teams in Cambridge include Google Cloud, Flights & Hotels, Search & Assistant, networking infrastructure, Android, YouTube, and Chrome. We also have employees working on research in areas of machine learning and artificial intelligence.

Other Google Incentives for this program
Google Value-Added Offerings

- Fixed Price with Google's Public Sector Subscription Agreement

  To address the challenge that public entities face of having pre-defined budgets that cannot be exceeded for a project, we are pleased to offer the Public Sector Subscription Agreement (PSSA). This unique offering from Google was created to address these challenges by enabling our state and local government and education customers to purchase Google Cloud services at a fixed monthly price for a defined workload, regardless of actual consumption. This allows public sector customers, such as universities and government agencies, to achieve predictable costs, eliminate queues for resources, and easily scale computing power for large projects.

  The PSSA provides access to Google Cloud's full suite of services, including computing, storage, networking, security, data analytics, and advanced tools in Artificial Intelligence (AI) and Machine Learning (ML).

- Google Public Sector - Rapid Innovation Team (RIT)

  This team is made of industry executives and technologists and they focus on identifying challenges that government customers face in their job and prototype solutions to address these challenges. Their mission is to bring the 'Google Culture of Innovation' to our customers through proving mission challenges are possible and packaging them to scale 10x. We define "Rapid" as delivery of a high-fidelity prototype within 6-8 weeks.

  Google would welcome the opportunity to collaborate with the Massachusetts Technology Collaborative on challenges they are facing and offer Google's point of view for how we might be able to assist.

- No Cost Training for Massachusetts Technology Collaborative

  Google provides comprehensive resources for career advancement through its Google Cloud Skills Boost and Grow with Google Career Certificates platforms.

Google Cloud Skills Boost enables individuals and teams to deepen their expertise in Google Cloud technologies, offering specialized training, skill badges, and industry-recognized certifications for roles like cloud architects, data engineers, and security specialists.

In parallel, Grow with Google Career Certificates delivers accessible, practical training in high-demand fields such as IT support, data analytics, project management, and UX design, designed for learners with no prior experience or traditional degrees. Both platforms equip users with the skills needed to thrive in job roles relevant to government agencies and educational institutions, fostering professional growth and career success.

- Explore Cloud Skills Boost
- Earn a Grow with Google certificate

III. Proposed Solution Architecture: The Integrated Data Commons

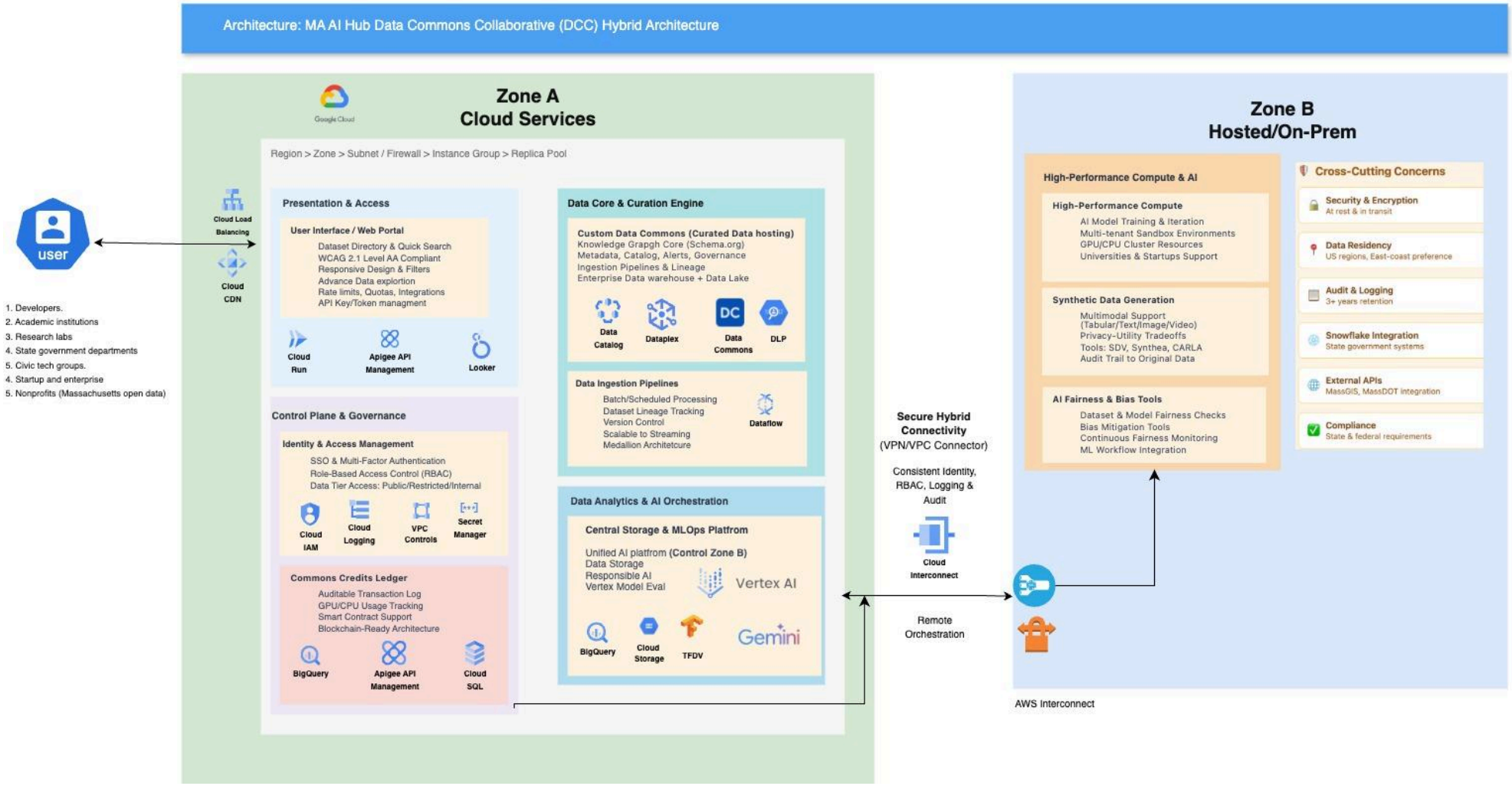
The CLOUDSUFI solution transforms the DCC from a static data repository into a **dynamic, interconnected living knowledge graph** for the Commonwealth. This technical architecture is based on the **proven, documented Data Commons reference architecture**. In addition to traditional catalog and query tools, the platform incorporates **semantic search and natural language (NLP) interfaces** that allow users to interact with the Data Commons using everyday language. Rather than requiring knowledge of table names or schemas, researchers, policymakers, and innovators can simply pose questions in natural language. Behind the scenes, the system applies **language models, embeddings, and ontology mapping** to translate those questions into structured queries and align terminology across datasets. These capabilities ensure the DCC is not only a secure and scalable repository, but also an **intuitive and inclusive discovery environment**, lowering technical barriers and accelerating insight generation across the Commonwealth's research and innovation ecosystem.

1. **Core Technology (Knowledge Graph):** The foundation is the Data Commons knowledge graph, which employs a standardized, open schema derived from Schema.org. Data is structured as semantic triples (entities, properties, relationships), allowing disparate datasets (e.g., census, economic, and health data) to be mapped to a single, canonical node (e.g., "Suffolk County"). This architecture facilitates immediate **cross-domain analysis** and eliminates time-consuming "data wrangling".

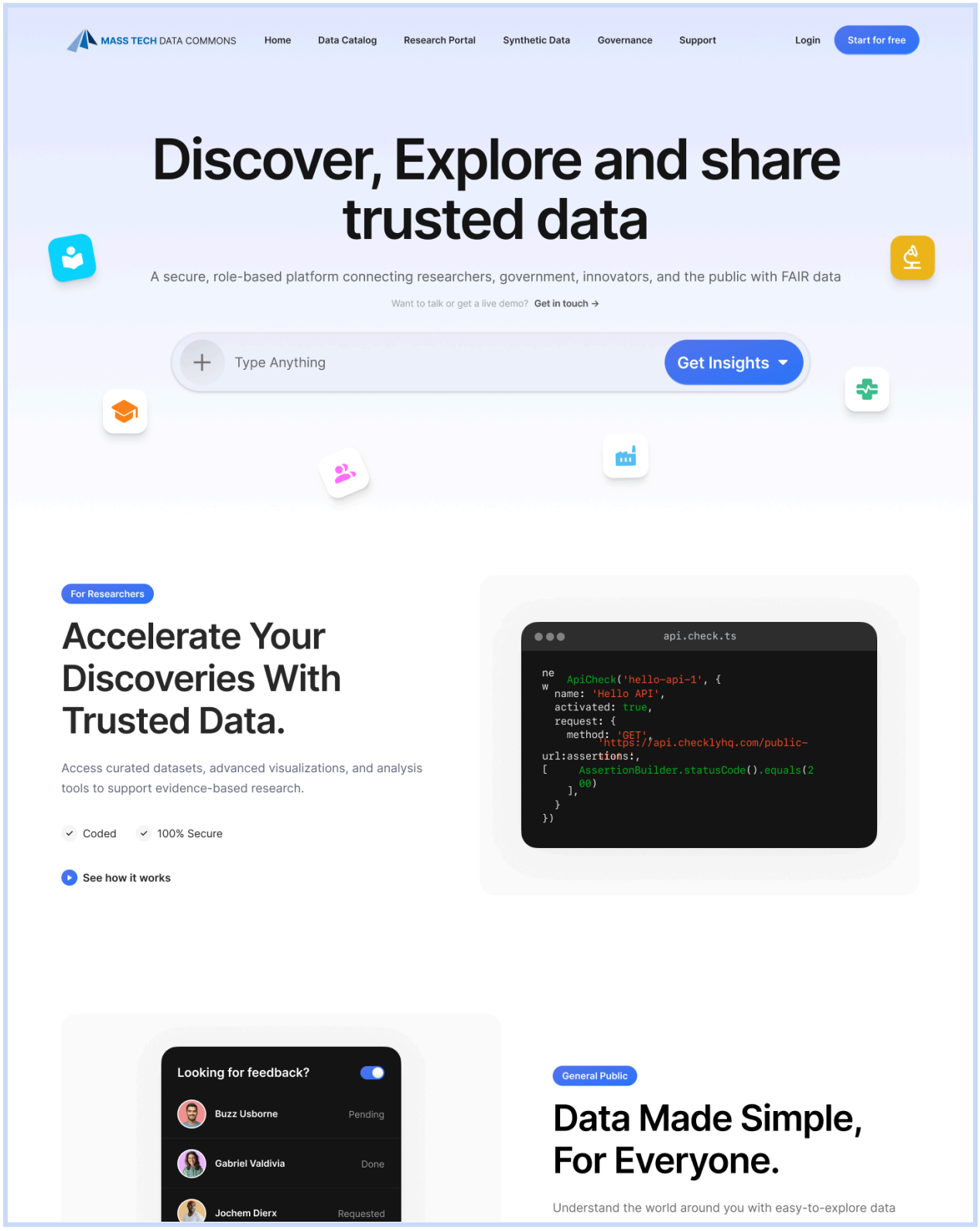
2. **FAIR Compliance:** The knowledge graph design is the core implementation of **FAIR principles** (Findable, Accessible, Interoperable, Reusable). The unified schema makes data inherently Interoperable, and its APIs ensure high Accessibility.

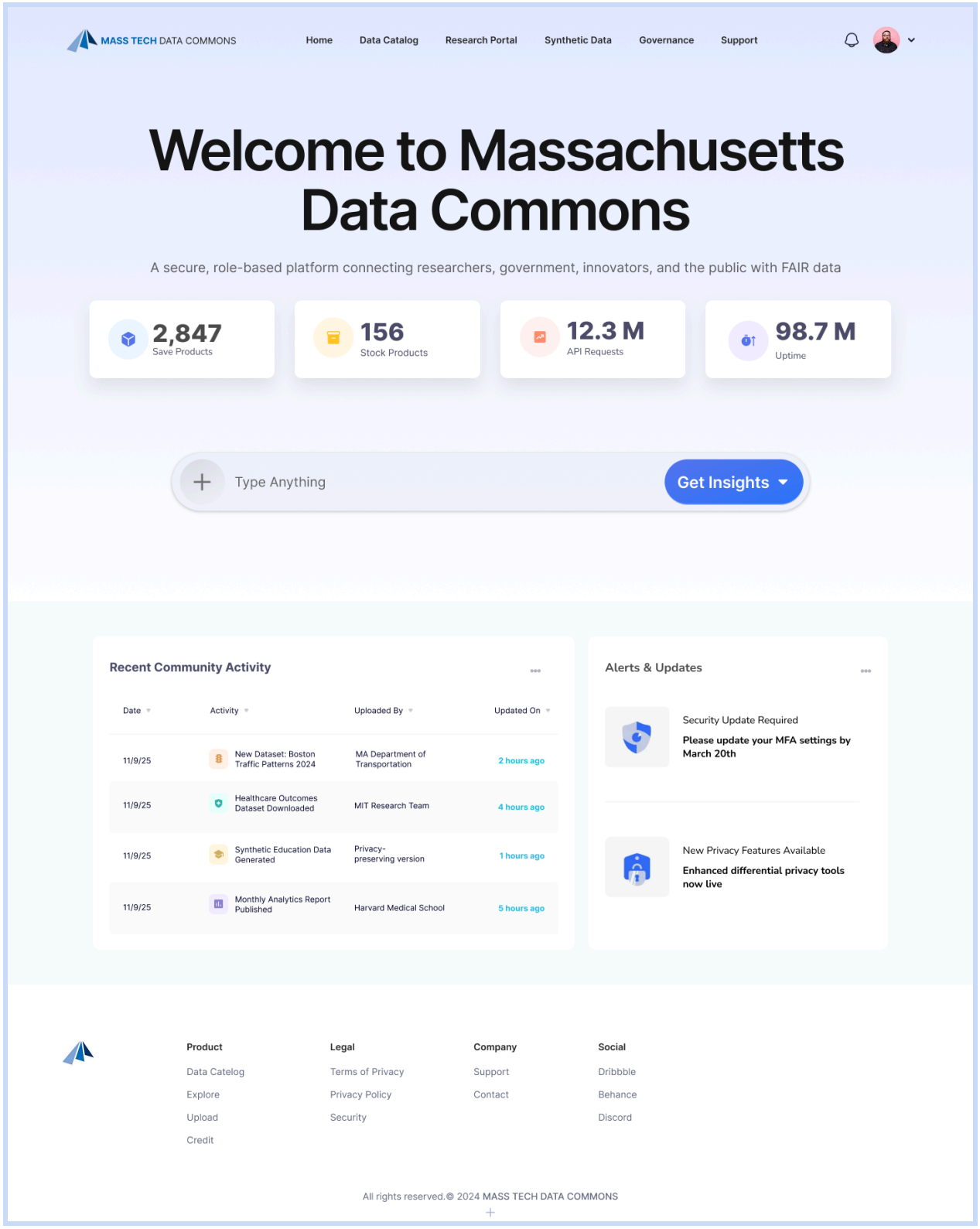
3. **Hybrid and Scalable Deployment:** The system is designed for a **hybrid deployment model**. The architecture will leverage **MGHPCC/AICR** for hosted data and compute, paired with **Cloud-based services** (GCP) for cataloging, APIs, and web access, ensuring consistent identity, RBAC, logging, and auditing across both domains. All data processing and storage will be in **U.S. regions** with an East-coast preference. The architecture is modular, API-first, and standards-aligned to enable seamless future expansion (e.g., multi-cloud federation, quantum computing integration).

High Level Architecture below


Architecture: MA AI Hub Data Commons Collaborative (DCC) Hybrid Architecture
Zone A
Cloud Services
Region > Zone > Subnet / Firewall > Instance Group > Replica Pool
user
1. Developers.
2. Academic institutions
3. Research labs
4. State government departments
5. Civic tech groups.
4. Startup and enterprise
5. Nonprofits (Massachusetts open data)
Cloud Load Balancing
Cloud CDN
Presentation & Access
User Interface / Web Portal
Dataset Directory & Quick Search
WCAG 2.1 Level AA Compliant
Responsive Design & Filters
Advance Data explortion
Rate limits, Quotas, Integrations
API Key/Token managment
Cloud Run
Apigee API Management
Looker
Control Plane & Governance
Identity & Access Management
SSO & Multi-Factor Authentication
Role-Based Access Control (RBAC)
Data Tier Access: Public/Restricted/Internal
Cloud IAM
Cloud Logging
VPC Controls
Secret Manager
Commons Credits Ledger
Auditable Transaction Log
GPU/CPU Usage Tracking
Smart Contract Support
Blockchain-Ready Architecture
BigQuery
Apigee API Management
Cloud SQL
Data Core & Curation Engine
Custom Data Commons (Curated Data hosting)
Knowledge Grapgh Core (Schema.org)
Metadata, Catalog, Alerts, Governance
Ingestion Pipelines & Lineage
Enterprise Data warehouse + Data Lake
Data Catalog
Dataplex
Data Commons
DLP
Data Ingestion Pipelines
Batch/Scheduled Processing
Dataset Lineage Tracking
Version Control
Scalable to Streaming
Medallion Architetcure
Dataflow
Data Analytics & AI Orchestration
Central Storage & MLOps Platfrom
Unified AI platfrom (Control Zone B)
Data Storage
Responsible AI
Vertex Model Eval
Vertex AI
BigQuery
Cloud Storage
TFDV
Gemini
Secure Hybrid Connectivity
(VPN/VPC Connector)
Consistent Identity, RBAC, Logging & Audit
Cloud Interconnect
Remote Orchestration
Zone B
Hosted/On-Prem
High-Performance Compute & AI
High-Performance Compute
AI Model Training & Iteration
Multi-tenant Sandbox Environments
GPU/CPU Cluster Resources
Universities & Startups Support
Synthetic Data Generation
Multimodal Support (Tabular/Text/Image/Video)
Privacy-Utility Tradeoffs
Tools: SDV, Synthea, CARLA
Audit Trail to Original Data
AI Fairness & Bias Tools
Dataset & Model Fairness Checks
Bias Mitigation Tools
Continuous Fairness Monitoring
ML Workflow Integration
Cross-Cutting Concerns
Security & Encryption
At rest & in transit
Data Residency
US regions, East-coast preference
Audit & Logging
3+ years retention
Snowflake Integration
State government systems
External APIs
MassGIS, MassDOT integration
Compliance
State & federal requirements
AWS Interconnect

Sample Prototype of **Data Commons landing page for MASSTech**




MASS TECH DATA COMMONS
Home
Data Catalog
Research Portal
Synthetic Data
Governance
Support
Welcome to Massachusetts Data Commons
A secure, role-based platform connecting researchers, government, innovators, and the public with FAIR data
2,847
Save Products
156
Stock Products
12.3 M
API Requests
98.7 M
Uptime
Type Anything
Get Insights
Recent Community Activity
Date
Activity
Uploaded By
Updated On
11/9/25
New Dataset: Boston Traffic Patterns 2024
MA Department of Transportation
2 hours ago
11/9/25
Healthcare Outcomes Dataset Downloaded
MIT Research Team
4 hours ago
11/9/25
Synthetic Education Data Generated
Privacy-preserving version
1 hours ago
11/9/25
Monthly Analytics Report Published
Harvard Medical School
5 hours ago
Alerts & Updates
Security Update Required
Please update your MFA settings by March 20th
New Privacy Features Available
Enhanced differential privacy tools now live
Product
Data Catelog
Explore
Upload
Credit
Legal
Terms of Privacy
Privacy Policy
Security
Company
Support
Contact
Social
Dribbble
Behance
Discord
All rights reserved.© 2024 MASS TECH DATA COMMONS

IV. Technical Capabilities: Core Infrastructure & Cataloging

The core infrastructure is fully managed and serverless, optimizing performance and cost.

COMPONENT	TECHNOLOGY	RATIONALE AND CAPABILITY DETAIL
Data Lake/Staging	Google Cloud Storage (GCS)	Highly durable, cost-effective object storage for raw datasets.
Analytical Engine	Google BigQuery	Fully managed, petabyte-scale data warehouse with a serverless architecture, separating storage and compute for performance and cost control.
Backend Compute/API	Cloud Run	Fully managed container platform for Data Commons application logic (APIs, web front-end). Automatically scales down (even to zero) based on traffic, ensuring cost-effectiveness.
Transactional DB	Cloud SQL	Managed relational database service for application data and transactional integrity.
Data Cataloging	Google Cloud Data Catalog	**Superior Solution:** Replaces standalone catalogs (CKAN/Dataverse) with deep cloud integration. Provides **automated metadata ingestion** from BigQuery and GCS. Uses **schematized tag templates** for structured business metadata and integrates seamlessly with **Cloud DLP** for automated identification and tagging of PII/sensitive data.
Discovery/ Analytics	Looker with Gemini	Looker's semantic modeling layer (LookML) ensures a **single source of truth** for metrics. **Gemini in Looker** enables **AI-powered analytics** via natural language querying, accelerating insight generation for non-technical policy

		analysts and researchers. **AI-driven discovery** (e.g., semantic search) is supported.
Lineage/ Versioning	Data Commons Import Tools	Provenance tagging is included. At launch, **dataset-level lineage tracking** is provided, with architecture designed for eventual **column-level lineage** expansion.
Google Cloud Dataplex	Centralize data governance, improve data quality	an intelligent data fabric that provides a unified, serverless, and automated approach to manage and govern data across an organization's distributed data lakes, data warehouses, and data marts. It helps data scientists and analysts to find and use high-quality data more easily

The architecture is designed as a **secure, hybrid, and integrated platform** that separates core cloud services (Zone A) from high-performance, potentially sensitive workloads (Zone B Hosted/On-Prem). This structure ensures governance, scalability, and adherence to the strict security and ethical requirements of the RFP.

The architecture is segmented into five primary functional layers, linked by secure connectivity and governed by a central control plane.

1. Presentation and Access Layer

This layer handles all user interaction, ensuring ease of use and compliance with accessibility standards.

- **User Interface:** A central **User Interface / Web Portal** is provided for interaction. It must be **WCAG 2.1 Level AA Compliant** and offer **Responsive Design**.
- **Discovery:** This portal provides a **Dataset Directory & Quick Search** capability, along with **Advanced Data Exploration** features, including semantic/NLP searches.
- **API Management:** Access to data and services is governed by **Apigee API Management**, which provides a secure and scalable gateway for all API interactions. Apigee manages **API keys/tokens**, ensuring only authorized access, and enforces **rate limits and quotas** to prevent abuse and maintain fair resource allocation across tenants. These policies can be configured at multiple levels (per API product, per developer app, or per user) and monitored in real time. In addition, Apigee

supports **traffic shaping, analytics, and automated quota resets**, giving administrators granular control over usage while ensuring reliable performance for all participants. Beyond access control, Apigee enables seamless **integrations** with internal and external systems, including support for versioning, mediation, and lifecycle management of APIs.

- **Connectivity:** Access for users (Developers, Academic institutions, Government departments, etc.) is secured via **Cloud Load Balancing** and **Cloud CDN**. Secure access to on-prem resources is managed via **Secure Hybrid Connectivity (VPN/VPC Connector)** and potentially **Cloud Interconnect** for high bandwidth.
- **API Usage Policies (initial defaults, tunable):**

  • **Burst control:** short-window throttles to protect shared services during spikes.
  • **Sustained rate-limits:** per API product and per developer app (e.g., per-minute ceilings).
  • **Daily quotas with automated reset** and **exception workflows** for approved research bursts.
  • **Tenant-aware keys/tokens** to ensure fair allocation across universities, startups, agencies.
  • **Real-time monitoring & alerts** (traffic, errors, latency) with audit trails of policy changes.

2. Control Plane & Governance

This is the non-negotiable security and audit layer that ensures compliance with the strict governance requirements (e.g., RBAC, NIST alignment) previously discussed.

- **Identity Management: Identity & Access Management (IAM)** enforces **SSO & Multi-Factor Authentication** (MFA). **Cloud IAM** implements **Role-Based Access Control (RBAC)** to manage access across the required data tiers: **Public, Restricted, and Internal**.
- **Security Perimeter:** The platform is secured using **VPC Controls** (VPC Service Controls perimeter) and the **Secret Manager** to protect sensitive credentials.
- **Auditing and Metrics: Cloud Logging** provides the mandatory audit functionality, ensuring **Consistent Identity, RBAC, Logging & Audit** across the system. A unique feature is the **Commons Credits Ledger**, which provides an **Auditable Transaction Log** for tracking **GPU/CPU Usage** and is architected to be **Blockchain-Ready** to support smart contract functionality.

3. Data Core & Curation Engine

This foundational layer is responsible for ingestion, metadata management, and structuring the data into a usable ecosystem. This addresses the RFP requirement for strong metadata management and moving beyond simple catalogs.

- **Knowledge Graph Core:** The data organization is centered on a **Knowledge Graph Core** that utilizes **Schema.org**, distinguishing it from older, dataset-centric catalogs (like CKAN/Dataverse) [172, 8.10, 55]. This enables inherent interoperability (FAIR principles).
- **Metadata Management: Data Catalog** is used for managing Metadata, Catalog functions, Alerts, and Governance.
- **Ingestion and Lineage: Data Ingestion Pipelines** utilize **Dataflow** for Batch/Scheduled Processing. The architecture supports the required **Dataset Lineage Tracking** and **Version Control** capabilities. The storage architecture employs a **Medallion Architecture** (layers of data quality).
- **Storage and Security:** Core storage uses **BigQuery** (Enterprise Data Warehouse) and **Cloud Storage** (Data Lake). **DLP (Data Loss Prevention)** is integrated to scan and protect sensitive data during ingestion and use.
- **Interoperability Standards:** The ingestion and modeling layer supports **healthcare interoperability** (e.g., **HL7/FHIR** schemas and vocabularies) and aligns with **W3C/Linked-Data** practices (e.g., **DCAT**, **Schema.org**) to advance FAIR principles across domains.

4. Central Storage & MLOps Platform (Zone A)

This layer provides the primary compute and storage resources for analytical workloads and serves as the command center for the hybrid architecture.

- **AI/ML Orchestration: Vertex** serves as the **Unified AI Platform**, responsible for the orchestration and control of resources in Zone B. It integrates with **BigQuery** and **Cloud SQL/Firestore (Commons Credits)**.
- **Responsible AI:** The platform integrates **Responsible AI** components, including **Vertex Model Eval**, which directly supports the RFP requirements for model fairness checks and continuous monitoring (as discussed in the history).

5. Zone B: Hosted/On-Prem - High-Performance Compute & AI

Zone B is architected as a highly secure, segregated environment focused on demanding workloads, aligning with the RFP's need for advanced, privacy-sensitive AI capabilities.

- **High-Performance Compute (HPC):** This zone hosts the **High-Performance Compute** resources, including **GPU/CPU Cluster Resources** in **Multi-tenant Sandbox Environments**. This is designed specifically to support external partners (Universities & Startups).
- **Synthetic Data Generation:** This critical RFP requirement is housed here. It supports **Multimodal Support** (Tabular/Text/Image/Video) and allows for fine-tuning **Privacy-Utility Tradeoffs**. It specifically employs leading open-source tools cited in the RFP, such as **SDV, Synthea, and CARLA**. Crucially, it must provide an **Audit Trail to Original Data** for compliance.
- **AI Fairness Tools:** Zone B hosts the specialized tools required for ethical AI development, including **Dataset & Model Fairness Checks** and **Bias Mitigation Tools**. These capabilities ensure **Continuous Fairness Monitoring** integrated into the ML Workflow, fulfilling the RFP requirements for fairness tools (like AIF360/Fairlearn).
- **Hardware & Hosting Profile (Zone B):** The DCC's high-performance workloads run at **MGHPCC/AICR** in a **secure, multi-tenant cluster** designed for elastic growth. The environment provides:
    - **GPU/CPU compute pools** for training/inference and synthetic data generation, provisioned via cluster schedulers with **per-tenant quotas** and isolation.
    - **High-throughput parallel storage** for short-lived training data and artifacts, plus policy-driven replication to long-term stores.
    - **Secure hybrid connectivity** (private links) to Zone A services for identity, catalog, logging, and credits metering.
    - **Initial capacity posture:** Sized for **hundreds of concurrent users** at launch with a roadmap to scale to **1,000+ concurrent users** through additional node pools and storage tiers; capacity reviews are part of quarterly governance.
    - **Operations note:** Hardware SKUs and final counts will be confirmed with **MGHPCC/AICR** and MassTech; the design supports gradual expansion without architectural change.

TOOL/SERVICE	LAYER	FUNCTION IN ARCHITECTURE	REASONING/RFP ALIGNMENT
BigQuery / Cloud Storage	Data Core & Central Storage	Provides the scalable **Enterprise Data Warehouse** and **Data Lake** infrastructure.	Essential for storing vast amounts of structured and unstructured data, supporting the required Medallion Architecture and data analytics workloads.
Data Catalog	Data Core	Centralized system for managing **Metadata, Catalog, Alerts, and Governance**.	Crucial for implementing metadata standards and achieving the **Findable** aspect of FAIR data principles. It forms the structure for the **Custom Data Commons**.
Dataflow	Data Core	Manages **Batch/Scheduled Processing** pipelines and ensures scalability to **Streaming** data.	Provides the scalable infrastructure required for the **Data Ingestion Pipelines**, supporting automated processing and ensuring data quality.
DLP (Data Loss Prevention)	Data Core	Used for integrating security and data scanning into the core pipeline.	Required to identify and protect sensitive data to comply with security standards (e.g., HIPAA, MA state privacy laws, as noted in the conversation history).
Cloud IAM	Control Plane & Governance	Implements **SSO, MFA**, and **Role-Based Access Control (RBAC)**.	Fulfills the foundational RFP requirement for robust **Security/Governance Frameworks**, ensuring data access is managed based on the user's role and data tier (Public/Restricted/Internal).

VPC Controls	Control Plane & Governance	Establishes a security perimeter around cloud services.	Crucial for preventing data exfiltration and maintaining the required level of security and boundary controls mandated by governance standards.
Apigee API Management	Presentation & Access	Manages APIs, **API Key/Token management**, and enforces **Rate limits**.	Provides secure external access, controlling and metering usage, which is necessary for managing resource consumption within the Commons Credits Ledger system.
Vertex (AI Platform) / Model Eval	MLOps & Central Storage	Acts as the **Unified AI Platform** (MLOps hub) and provides tools for model evaluation.	Provides the orchestration capability required to manage model training (controlling Zone B) and integrates essential **Responsible AI** tools, supporting **AI fairness** checks.
SDV, Synthea, CARLA	Zone B HPC	Specific tools for **Synthetic Data Generation**.	These tools are explicitly required (as illustrative examples) by the RFP to provide a robust, multimodal synthetic data capability, managing **Privacy-Utility Tradeoffs** in a dedicated secure zone.
TFDV (TensorFlow Data Validation)	Data Analytics & AI Orchestration	Likely used for proactive data analysis and validation within the pipeline.	Essential for pre-processing checks and detecting bias or anomalies early in the data preparation phase, supporting the required AI Fairness measures (similar to AIF360/Fairlearn referenced in history).

V. Enabling Financial Governance: The Commons Credits System

The DCC requires a central ledger to track "Commons Credits" to incentivize contributions, promote efficient usage, and sustain ecosystem engagement. This is implemented as a closed-loop incentive system, not a real-currency mechanism at launch, but can be expanded in future phases to include actual currency. Please refer to the below figure for a high level architecture for the commons credits approach.

Commons Credits Implementation Approach

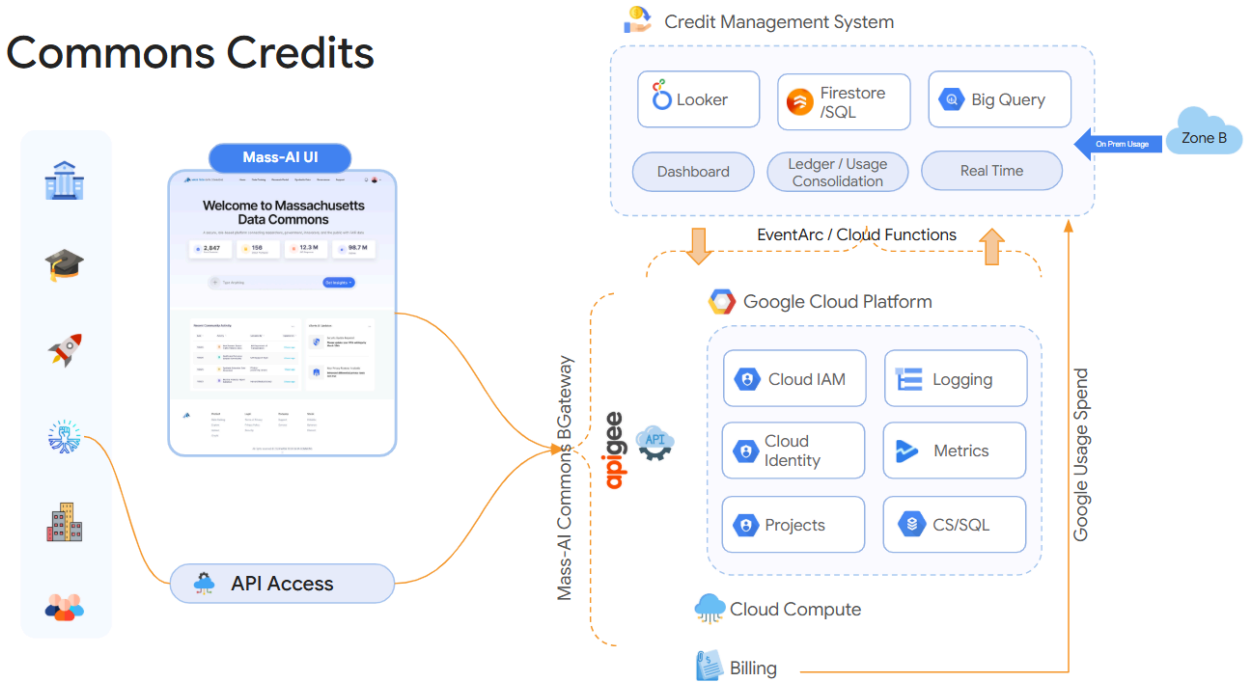
1. Unified API Access
    a. All usage flows through a single API layer managed by Apigee.
    b. The Mass-AI UI and programmatic clients consume the same APIs, ensuring consistent metering and policy enforcement.
2. Credit Services on Cloud Compute
    a. A dedicated Credit Management System runs as microservices on Cloud Run.
    b. Core endpoints include authorize, consume, refund, get-balance, and top-up.
3. Event Capture with Eventarc/Cloud Functions
    a. Every API call or backend job generates a usage event containing tenant, resource, and unit details.
    b. Eventarc or Cloud Functions routes these events to the credit processor for handling.
4. Ledger Implementation
    a. The **Cloud SQL database** serves as the **transactional system of record**, ensuring atomicity for balance changes and providing an auditable transaction history.
    b. All usage events, top-ups, refunds, and adjustments are also appended to **BigQuery** as an **immutable ledger**, consolidating internal Google Cloud usage (via Billing export) and external/on-prem usage into a single source of truth.
    c. The platform is designed to support **smart contracts** for automated enforcement of credit rules, and includes a **pluggable adapter layer** to support future blockchain integration without re-architecting. This balances audit integrity, usability, and flexibility.
5. Real-Time Balance in Firestore/SQL
    a. Current balances and entitlements are stored in Firestore or Cloud SQL for low-latency checks.

   b. The credit processor updates this store atomically, ensuring accurate real-time enforcement at request time.

6. Tracking & Reporting
   a. Credit usage logs (including API calls, data downloads, and compute resource consumption such as GPU/CPU usage) are streamed to **BigQuery**.
   b. **Looker dashboards** provide administrators and tenants with visibility into balances, burn-rates, usage patterns, and forecasts, supporting transparency and oversight needs.
   c. Batch reporting is sufficient initially, with options to evolve toward near real-time dashboards as adoption grows.
7. Billing and Reconciliation
   a. Google Cloud Billing data is exported to BigQuery and reconciled with the credits ledger.
   b. This enables cost recovery, grant allocation, anomaly detection, and margin analysis.
8. Payments and Top-Ups
   a. An **external payment processor** (e.g., Stripe, Adyen, or Kill Bill) handles monetary transactions.
   b. Successful payments trigger **TopUp events** that are recorded in BigQuery and reflected in Cloud SQL/Firestore balances.
   c. Peer-to-peer trading and real-currency purchases are **not in scope initially**.

9. Access and Integrity
   a. The system leverages **Cloud IAM** for secure authentication, with MFA required and role-based access control (RBAC) to protect sensitive operations.
   b. Cloud Audit Logs track all administrative changes, and immutable BigQuery facts plus WORM copies in Cloud Storage ensure tamper-evident integrity. The logs will act as a catch-all enabler for any custom events that need to be used for common credits computations.

10. Controls and Governance
    a. Quotas and rate limits are enforced at the API gateway.
    b. All financial events include idempotency keys to prevent duplication.
    c. Nightly recomputation of balances from the BigQuery ledger corrects drift and strengthens audit confidence.



Commons Credits High Level Architecture

VI. Differentiating AI Capabilities

Our integrated approach leverages Vertex AI to deliver superior capabilities for both synthetic data generation and ethical AI governance, which are critical components of the MA AI Hub vision.

VI.A. Synthetic Data Generation (RFP 2.4)

CLOUDSUFI moves beyond integrating standalone, domain-specific tools (SDV, Synthea, CARLA) by providing a **unified, multi-modal generative platform** centered on **Vertex AI**.

- **Technology: Gemini models** running on Vertex AI will generate high-quality, privacy-preserving synthetic data across diverse formats and modalities (including text, tabular, and multimodal data).
- **Workflow Integration:** The synthetic data pipeline is managed entirely within the **secure, governed Google Cloud environment** (BigQuery, Vertex AI Training). This eliminates risky data movement and simplifies MLOps.
- **Traceability and Validation:** The system supports schema-driven generation (e.g., using Snowfakery) and is required to maintain **audit trails linking synthetic datasets back to original real datasets** and validate configurable privacy-utility tradeoffs.
- **Target Domains:** Priority domains for synthetic data pilots include **lifesciences, robotics, education, and healthcare**.

VI.B. Responsible AI and Bias Framework (RFP 2.5)

We deploy **Google's comprehensive Responsible AI (RAI) toolkit**, embedding fairness, interpretability, and safety across the entire machine learning lifecycle.

- **Bias Detection in Data: TensorFlow Data Validation (TFDV)** will proactively profile training datasets to identify potential bias sources (e.g., data skew or under-representation) across protected attributes.
- **Model Evaluation and Debugging:** Post-training models will be evaluated using **Fairness Indicators** and the interactive **What-If Tool** to inspect performance metrics (e.g., accuracy, false positive rates) across sensitive features (e.g., race, gender, geography, socioeconomic factors).
- **Operationalized Monitoring: Vertex AI Model Evaluation** provides built-in data and model bias metrics, making fairness evaluation a first-class citizen in the MLOps pipeline. **Continuous monitoring** of fairness drift in deployed models is required.
- **Virtuous Cycle:** The integrated platform connects the synthetic data and fairness capabilities, allowing for the generation of synthetic data to augment under-represented minority groups and subsequent validation using TFDV and Fairness Indicators, all within a single, secure environment.
- **Reporting:** Fairness reporting is primarily for **internal auditing and compliance** in this phase, though it must enable public-facing transparency features in the future.

VI.C Deeply Embedded in Google Cloud's Innovation Efforts

CloudSufi is not an outside implementer but a trusted collaborator helping shape Google Cloud's native platforms. We are:

- Actively co-developing a governance UI layer on top of Google Cloud Dataplex, integrating metadata, lineage, and quality APIs for real-time visibility and policy enforcement.
- Engaged in the multi-year Data Commons program, having been awarded the three-year Data Commons Data Loading Statement of Work (SOW) — underscoring our role in expanding and operationalizing Google's Data Commons knowledge graph.
- Focused on extensibility and scale, ensuring that Massachusetts benefits not only from proven Google technologies (Data Commons, Dataplex, BigQuery, Apigee, Vertex AI, Looker) but also from CloudSufi's direct, ongoing contributions to these ecosystems.

This close alignment with Google Cloud means the Massachusetts AI Hub is not adopting experimental technology or unproven tools; it is leveraging the same operational infrastructure and expertise that powers Google's innovation roadmap.

Key Scope & Accomplishments

- Designing and developing a user-friendly governance UI layer on top of Google Cloud Dataplex.
- Integrating Dataplex metadata, lineage, and data quality APIs for real-time visibility into data assets.
- Enabling policy management, data cataloging, and compliance workflows directly within the UI.
- Automating metadata synchronization between Dataplex and the governance platform.
- Supporting governance at scale across multiple domains and distributed environments.
- Delivering a scalable, extensible solution adaptable to evolving enterprise needs.

Business Benefits

- Simplified Data Governance – intuitive interface for policy management and compliance.
- Improved Visibility – clear insights into data lineage, quality, and asset health.
- Enhanced Collaboration – enables stakeholders to govern and share data seamlessly.
- Faster Decision-Making – accelerates access to trusted, governed data.
- Scalability – designed for large, distributed data environments without complexity.

VII. Security, Governance, and Compliance Framework

Security is implemented through a multi-layered, defense-in-depth posture native to GCP, ensuring compliance with state requirements. **Audit Retention & Archival: ≥3 years** log retention at launch with an **archival strategy to be finalized** with MassTech. Logs are exportable to SIEM (e.g., Splunk/ELK) for security operations and compliance reporting

SECURITY & GOVERNANCE REQUIREMENT	CLOUDSUFI SOLUTION (GCP COMPONENT)	REQUIREMENT DETAIL/CITATION
Access Control & Identity	Cloud IAM / Identity Platform	Granular **Role-Based Access Control (RBAC)** and support for **Multi-Factor Authentication (MFA)** are mandatory. Supports multi-tiered access: Public, Restricted (requires review), and Internal.
Dataset Tiering & Security	Cloud Data Catalog Policy Tags	Enables fine-grained, **column-level security** within BigQuery to control access to sensitive information. Adheres to data classification security standards set by the MA CISO.
Data Protection	Default Encryption (CMEK Option)	All data is **encrypted by default** at rest (in GCS/BigQuery) and in transit. Supports the use of Customer-Managed Encryption Keys (CMEK) for Commonwealth key authority.
Network Perimeter	VPC Service Controls	Establishes a virtual security boundary, preventing data exfiltration even if credentials are compromised.
Auditing and Logging	Cloud Audit Logs / SIEM Integration	Provides a complete, immutable record of all administrative actions and data access events. Audit logs must be retained for a minimum of **3+ years**. The system is designed for integration with SIEM tools (Splunk, ELK).

Data Residency	US Regions (East-coast Preference)	All data (including logs and model artifacts) will be stored and processed in **U.S. regions** with an East-coast preference. **Remote access** to Commonwealth Data from outside the continental United States is **prohibited**.
Consent Management	Cloud SQL / Integrated Workflow	Tokenized, revocable, and time-bound consent management is **in scope for Phase 1** for restricted datasets.
Governance Alignment	NIST/CISA/Applicable Laws	Governance adheres to applicable state and federal laws. Vendors are encouraged to align with recognized standards like the **NIST AI Risk Management Framework**. Defining governance policies will be a **collaborative effort** with MassTech.

VIII. Project Management, Delivery, and Ecosystem Growth

CLOUDSUFI (via Google Cloud PSO) will lead the implementation using a proven methodology, focusing on transparent program management, stakeholder coordination, and phased delivery. The anticipated start date is **December 2, 2025**. A **phased release** approach is preferred to achieve early usability.

1. Project Delivery & Environments

   We set up **Dev/Test, QA, and Production environments**. The delivery plan includes rigorous QA testing, tailored training programs for end-users and administrators, and comprehensive documentation.

2. User Experience & Accessibility

   The platform must meet **WCAG 2.1 Level AA accessibility standards** at a minimum. The architecture will support **multilingual accessibility** (e.g., screen reader compatibility for non-English content) for future localization.

   Accessibility QA: We will validate **WCAG 2.1 AA** with **screen reader compatibility**, **keyboard-only navigation**, and **captioning/transcripts** for multimedia. Testing includes assistive-technology checks and UAT with accessibility criteria, with results documented for acceptance.

3. Data Curation and Integration

   We will collaborate with MassTech to refine the initial dataset strategy and propose the initial set of datasets, drawing from the illustrative list (Appendix A). Integration will use **proven, open-source Data Commons import tools** that map local data into the knowledge graph's schema. We will leverage APIs for integration with external state systems (e.g., MassGIS, MassDOT).

4. Ecosystem Engagement (RFP 5.3)

   We are committed to fostering the local ecosystem.

   - **Academia:** Provide **Google Cloud research credits** and deliver workshops/training modeled on existing Data Commons educational courseware.
   - **Startups:** Offer Cloud credits, mentorship, and business support through the **Google for Startups Cloud Program**.
   - **Government:** Collaborate with EOTSS to integrate the DCC's queryable knowledge graph capabilities with the existing Massachusetts Data Hub.

Total Estimated Duration (Kickoff to Public Launch): **40 Weeks** (Accelerated Phased Release)

CLOUDSUFI proposes an **accelerated 40-week phased release** schedule, leveraging the inherent efficiencies and pre-built components of the **Google Data Commons architecture on GCP**. This approach directly addresses the RFP's preference for a phased release and feasibility of timeline.

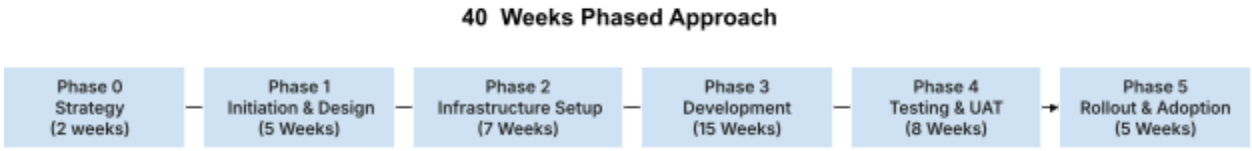
The plan is structured around the five mandated Program Phases outlined in DCC Requirements Section 3.0. We prioritize the implementation of the core foundation (Phase 2), including the integrated **Cloud Data Catalog** and the **Commons Credits ledger**, followed immediately by the deployment of differentiating AI capabilities (**Vertex AI for Synthetic Data** and the **Responsible AI (RAI) Toolkit**) in Phase 3. This ensures the platform is secure, functional, and equipped with high-value AI features for initial pilot users well ahead of the final public launch.

Key Acceleration Drivers

1. **Platform Instantiation:** We avoid custom-building the knowledge graph, instead instantiating a customized, secure version of the **Google Data Commons knowledge graph**, significantly accelerating the Data Integration phase.
2. **Native Integration:** Utilizing tightly integrated GCP services (BigQuery, Vertex AI, Cloud Data Catalog) eliminates time lost integrating fragmented open-source tools mentioned in the RFP.
3. **Concurrent Development:** We run website, metadata, and AI feature development concurrently within the mandated **Dev/Test and QA environments**.

Detailed Project Plan (40 Weeks)

The plan outlines phased delivery of MA AI Hub Data Commons Collaborative, through a phased and sprint-based delivery approach, ensuring alignment with business objectives, technical readiness, and governance requirements. The plan is structured into five phases, with activities running in parallel to optimize delivery speed and value realization:



Phase 0

Investment -Strategic Assessment (2 Weeks): CLOUDSUFI, supported by Google Cloud, will dedicate an expert Assessment POD to run a rapid 2-week discovery using our proven **Data Commons Readiness Framework**. This phase aligns key stakeholders, validates assumptions, and surfaces strategic opportunities before the build begins. Deliverables include a readiness baseline, risk map, and high-level blueprint that de-risks execution and accelerates decision-making. As an **investment** from our side, this ensures MassTech enters Phase 1 with clarity, consensus, and a focused roadmap

Phase 1

Initiation & Design (5 Weeks): Project kickoff with team alignment and charter approval. Finalize detailed architecture and governance strategy. Review and approve dataset curation criteria and confirm the initial dataset list.

Phase 2

Infrastructure & Setup (7 Weeks): Provision cloud infrastructure, identity management, and foundational services. Validate development and test environments to support synthetic tooling and backend deployment.

Phase 3

Development (15 Weeks): Complete initial data ingestion pipelines and metadata catalog. Operationalize the synthetic data engine to produce test outputs. Deliver a functional website with browsing, search, and dataset detail pages. Deliver an admin dashboard for dataset management.

Phase 4

Testing & Governance (8 Weeks): Conduct platform QA across functional, performance, and security dimensions. Implement the governance framework, including RBAC, data tiers, and usage agreements.

Phase 5

Rollout & Adoption (5 Weeks): Finalize user documentation and deliver tailored training. Onboard initial user groups. Gather pilot feedback and prepare the platform for public launch and ecosystem adoption.

This 2 weeks of sprint-driven approach ensures that platform infrastructure, AI capabilities, and governance frameworks evolve in tandem, supporting a robust, secure, and scalable delivery

SPRINT	WEEKS	KEY ACTIVITIES (PARALLEL WORKSTREAMS)	DELIVERABLES / OUTCOME	DCC REQUIREMENT ALIGNMENT
Sprint 1	W1–W2	Initiation: Project Kickoff, core team resourced and aligned.	Project Kickoff, Architecture & Governance Defined	Project Initiation & Governance Policies
Sprint 2	W3–W4	Design: Finalize high-level Architecture & Governance strategy, define initial governance policies (ethical use, data tiering).	Architecture & Governance Defined	Project Initiation & Governance Policies
Sprint 3	W5–W6	Design: Refine and gain approval for Dataset Curation Criteria, finalize proposed data sourcing strategy and data types (tabular, text, image, etc.).	Dataset Curation Criteria Approved	Dataset Curation & Sourcing
Sprint 4	W7–W8	Design: Validate and confirm Initial Dataset List for hosting and external linking.	Initial Dataset List Confirmed	Dataset Curation & Sourcing
Sprint 5	W9–W10	Infrastructure: Platform Infrastructure Provisioned (Production environment setup: GCS, BigQuery, VPC Service Controls perimeter).	Platform Infrastructure Provisioned	Secure GCP Foundation & Access Management
Sprint 6	W11–W12	Infrastructure: Configure Cloud Identity & Access Management (IAM), define RBAC roles, implement Cloud Audit Logs.	IAM Configured	Secure GCP Foundation & Access Management

Sprint 7	W13–W14	Infrastructure: Dev/Test Environments Ready, deploy Data Commons backend (Cloud Run, Cloud SQL), configure AI Compute Resources (AICR).	Dev/Test Environments Ready	Dev/Test Environment Setup & Data Ingestion
Sprint 8	W15–W16	Development: Build initial Data Ingestion Pipelines using Data Commons import tools (CSV/MCF mapping), support non-standard formats.	Data Integration Started	Dev/Test Environment Setup & Data Ingestion
Sprint 9	W17–W18	Infrastructure: Deploy Google Cloud Data Catalog, implement automated metadata ingestion from BigQuery and GCS. Governance: Commons Credits Ledger Setup (Phase 1), implement Column-Level Security using Data Catalog Policy Tags.	Metadata Catalog Setup, Commons Credits Phase 1	Metadata Management & Governance Tools
Sprint 10	W19–W20	Development: Finalize Initial Data Catalog and Metadata System, integrate Cloud DLP for automated PII tagging, begin semantic search functionality.	Metadata Catalog Functional	Data Access, Privacy, and Website Functionality
Sprint 11	W21–W22	Development: Synthetic Data Engine Operational – provision Vertex AI environment, configure Gemini models. Website: Core functional browsing, basic search, filtering (WCAG 2.1 AA compliant).	Synthetic Data Engine Setup, Website Core Functional	Data Access, Privacy, and Website Functionality

Sprint 12	W23–W24	Development: Generate synthetic data test outputs for priority domains, validate configurable privacy-utility tradeoffs. Website: Implement Dataset Detail Pages, integrate Gemini in Looker for natural language querying.	Synthetic Data Engine Operational, Website Enhancements	Synthetic Data & Website Enhancements
Sprint 13	W25–W26	AI Capabilities: AI Fairness Framework Integration – integrate TFDV for data bias profiling, configure model evaluation hooks. Admin: Develop Admin Dashboard UI for managing submissions and user suggestions, implement multi-tenant administration capabilities. Data Integration: Complete pipelines for initial hosted datasets, configure version control and provenance tagging.	AI Fairness Setup, Admin Dashboard Live, Data Integration Complete	AI Fairness, Admin Controls & Provenance
Sprint 14	W27–W28	API Development: Finalize API-based data access (REST/Python APIs), implement API Gateway (rate limiting, API keys). Commons Credits: Finalize Sample Commons Credits ledger output, dashboard integration.	APIs Functional, Credits System Finalized	API Access & Commons Credits Integration
Sprint 15	W29–W30	Development: Finalize all core Website UI/API features, data integration, AI fairness, and credits system. Testing: Prepare for Internal QA of the end-to-end system.	Dev Deliverables Complete, QA Ready	QA Preparation & System Readiness

Sprint 16	W31–W32	Testing: Internal QA Testing – end-to-end testing of core functionality, synthetic data pipeline, and API access. Infrastructure: Performance tuning and load testing.	Platform QA, Performance Tested	Internal QA & Performance Validation
Sprint 17	W33–W34	Security: Independent security review against MA CISO standards, address critical findings. Governance: Fairness & Bias Reporting Dashboard operational, Governance Framework operational (tokenized, revocable, time-bound consent for restricted datasets).	Security Review Complete, Governance Framework Operational	Security & Governance Enforcement
Sprint 18	W35–W36	Training & Rollout: Finalize training and launch materials, tailored training for admins, researchers, and non-technical users. User Onboarding: Deliver training to initial administrators and pilot user groups, initiate pilot feedback collection. Ethical AI Training & Certification: Establish role-based training (researchers, admins, developers) covering responsible data use, bias detection, and RAI practices, with completion tracking and attestation prior to elevated access.	Training Delivered, User Onboarding & Pilot Feedback	User Training & Pilot Engagement
Sprint 19	W37–W38	Documentation & Engagement: Finalize Detailed Platform Architecture Documentation, formalize planned industry engagement	Documentation & Engagement Complete, Platform Stabilized	Documentation & Industry Engagement

		(researchers/credits, startups, EOTSS/Mass Data Hub integration). Public Launch Prep: Prepare for public launch, system stabilization.		
Sprint 20	W39–W40	Rollout & Post-Launch Support: Public Launch of MA AI Hub Data Commons Collaborative, monitor adoption, address critical post-launch issues, finalize knowledge transfer to operations.	Platform Stabilized, Knowledge Transfer Complete	Public Launch & Post-Launch Support

Project Governance & Change Management

Effective governance and structured change management are crucial for a project's success. This approach ensures the project delivers its intended outcomes while remaining adaptable to new requirements and unforeseen challenges.

Governance Approach

A clear governance framework will guide strategic decisions and maintain accountability -

- **Project Steering Committee:** This committee, comprising client and delivery leadership, will provide high-level oversight, resolve major issues, and make strategic decisions to keep the project on track.
- **Project Management Office (PMO):** The PMO will handle the day-to-day management of the project, including monitoring risks, tracking progress against milestones, and ensuring all activities comply with governance policies.
- **Defined Roles and Responsibilities:** Roles and responsibilities will be clearly defined and documented for all project stakeholders to ensure transparent decision-making and clear accountability.

Change Management Process

A disciplined change management process will handle changes efficiently and with full transparency -

1. **Change Identification:** Any proposed changes to the project's scope, new requirements, or unexpected issues will be formally submitted through a **change request process**.
2. **Impact Assessment:** Each request will undergo a thorough assessment to evaluate its impact on the **project's scope, schedule, cost, resources, and risks**.
3. **Approval Workflow:** The Steering Committee will review each request and either approve or reject it. The client will be directly involved in the approval of all material changes.
4. **Controlled Implementation:** Once approved, the change will be integrated into the project's baseline plan. This includes updating deliverables, timelines, and communication plans to all relevant stakeholders.

The structured approach ensures the project stays aligned with the priorities, while providing the discipline and transparency needed to manage changes without disrupting overall delivery.

RFP Requirements Compliance Checklist (Phase Integration)

The following checklist confirms that all required services and deliverables (RFP Section 2) are addressed within the proposed 40-week phased plan.

RFP SERVICE AREA (RFP SEC. 2)	RFP DELIVERABLE (RFP SEC. 2.8)	PHASE WHERE CORE IMPLEMENTATION OCCURS	CLOUDSUFI SOLUTION COMPONENT
1. Program Management & Coordination	Detailed Project Plan and Timeline	Phase 1 (W1) & Phase 5 (W36)	PSO Methodology & Integrated Project Plan
2. Technical Architecture & Platform Development	Detailed Platform Architecture Documentation	Phase 1 (W5) & Phase 5 (W36)	GCP Serverless Architecture (BigQuery, Cloud Run, Cloud SQL)
	Initial Data Catalog and Metadata System	Phase 2 (W9) & Phase 3 (W16)	**Google Cloud Data Catalog** (Automated sync, PII detection)
	Sample Commons Credits ledger output	Phase 2 (W10) & Phase 3 (W28)	Cloud SQL Ledger, BigQuery/Looker Reporting
3. Dataset Curation and Integration	(Integrated into plan)	Phase 1 (W4) & Phase 3 (W26)	Data Commons Knowledge Graph & Import Tools
4. Synthetic Data Capability	Synthetic Data Tooling and Sample Outputs	Phase 3 (W18 - W19) & Phase 4 (W32)	**Vertex AI with Gemini Models**

5. AI Fairness and Bias Framework	Fairness and Bias Reporting Dashboard	Phase 3 (W25) & Phase 4 (W31)	Google **Responsible AI Toolkit** (TFDV, Fairness Indicators)
6. Security, Governance & Compliance	Security and Governance Implementation	Phase 2 (W6 - W12) & Phase 4 (W33)	Cloud IAM, **VPC Service Controls**, Cloud Audit Logs
7. Testing, Rollout, and Adoption	Training and Launch Materials	Phase 4 (W30) & Phase 5 (W34)	Rigorous QA, Tailored Training Programs
8. Deliverables	Pilot user feedback	Phase 5 (W38)	Structured feedback collection from MassTech-recruited users
	Explanation of planned industry engagement	Phase 5 (W40)	Plan for Researchers (credits), Startups (programs), EOTSS/Mass Data Hub

IX. RFP Requirements Compliance Checklist

This table maps the RFP requirements and deliverables (Section 2) to the proposed CLOUDSUFI (Google Cloud) solution components, confirming compliance and indicating areas where the proposed solution exceeds requirements.

RFP SECTION	REQUIREMENT/DELIVERABLE	PROPOSED CLOUDSUFI SOLUTION & APPROACH	MEETS/ EXCEEDS
2.1 Program Management & Coordination	Integrated project plan, regular meetings, milestone tracking	Google Cloud PSO methodology leveraging deep collaboration and transparent communication.	Exceeds
2.2 Technical Architecture & Platform Development	Design/build website and backend infrastructure	Serverless architecture on GCP using BigQuery, Cloud Run, Cloud SQL, and Cloud Storage, based on proven Data Commons reference architecture.	Exceeds
	Implement metadata management, cataloging, and discovery tools (e.g., CKAN, Dataverse)	**Google Cloud Data Catalog** with automated ingestion, rich structured tags, and Google-powered search.	Exceeds
	Develop role-based access and identity management system	**Google Cloud IAM** and Identity Platform providing robust RBAC, MFA, and delegated administration.	Meets
	Develop a central ledger for tracking "Commons Credits"	Complete system using **Cloud SQL** (transactional ledger), **Cloud IAM** (identity), and **BigQuery/Looker** (reporting), designed for smart contracts and future pluggable backends.	Exceeds
	Enable API-based data access and usage analytics	Core Data Commons REST and Python APIs enabled. Usage analytics captured in BigQuery and reported via Looker	Meets

		dashboards. Rate limiting and API key management implemented.	
2.3 Dataset Curation and Integration	Assist in curating initial datasets	Collaborative curation using proven, open-source Data Commons import tools for harmonization into the knowledge graph.	Meets
	Ensure metadata completeness and FAIR data principles	Knowledge graph architecture is a native, mature implementation of FAIR principles.	Exceeds
	Support ingestion pipelines, version control, and provenance tagging	Data Commons import process includes provenance tagging and dataset-level version control.	Meets
2.4 Synthetic Data Capability	Integrate synthetic data tools (e.g., SDV, Synthea, CARLA)	Integrated synthetic data generation using **Gemini models on the Vertex AI platform**, supporting multi-modal data types within a governed MLOps environment.	Exceeds
2.5 AI Fairness and Bias Framework	Implement bias detection, auditing, and mitigation tools (e.g., AIF360, Fairlearn)	**Google's comprehensive Responsible AI (RAI) toolkit**, including TFDV, Fairness Indicators, What-If Tool, and native Vertex AI Evaluation metrics for continuous monitoring.	Exceeds
2.6 Security, Governance & Compliance	Define and implement RBAC, MFA, and dataset tiering; ensure logging, encryption, and audit trails	Multi-layered security including Cloud IAM, VPC Service Controls, default encryption (CMEK option), Cloud Audit Logs (3+ year retention), and Phase 1 tokenized consent.	Exceeds
2.7 Testing, Rollout, and Adoption	Conduct QA, train end users/administrators, prepare	PSO-led delivery including QA, tailored training, and comprehensive documentation for technical and	Meets

	documentation, support public launch	non-technical users. Compliance with WCAG 2.1 AA.	
2.8 Deliverables	Detailed Platform Architecture Documentation	Full GCP architecture documentation provided (Deliverable).	Meets
	Fairness and Bias Reporting Dashboard	Dashboards built in Looker and Vertex AI, leveraging Fairness Indicators and model evaluation outputs for internal auditing.	Meets
	Explanation of planned industry engagement	Detailed plan for engaging researchers (Cloud Credits), startups (Google for Startups), and government agencies (EOTSS integration).	Meets

X. Assumptions

The CLOUDSUFI project plan and technical architecture are based on the following key assumptions regarding the execution environment, data availability, and governance structure:

1. Architectural and Technical Assumptions

- **Hybrid Deployment Model Confirmed:** The solution will utilize a **hybrid model**, leveraging **MGHPCC/AICR** for hosted data and compute, paired with **cloud-based services** (e.g., Google Cloud Platform, as proposed by CLOUDSUFI) for the catalog, APIs, and web access.
- **API Availability for Integration:** Vendors can assume **APIs will be made available** for integration with external Massachusetts systems, such as MassGIS, MassDOT, and Public Health data sources.
- **Infrastructure Procurement:** MassTech will be responsible for **procuring the selected infrastructure independently**, including necessary licenses and hosting platforms; these are **not pass-through costs** for the vendor.
- **Environment Requirements:** Three distinct environments are required for delivery: **Dev/Test, QA, and Production**.
- **Data Ingestion Approach: Batch processing will be sufficient** for data ingestion in Phase 1, with real-time or streaming capabilities deferred to future phases.
- **Data Volume Estimates:** Initial data volumes are anticipated to be in the **tens of TBs with approximately 30% annual growth**.

2. Data and Scope Assumptions

- **Initial Dataset Scope:** The datasets referenced in DCC Requirements Appendix A are **illustrative examples and are not mandatory inclusions**; scope includes finalizing a recommendation for # datasets should be sourced & integrated during the initial launch.
- **Curation Responsibility:** The respondent is expected to **identify and propose high-value datasets** for inclusion, working in partnership with MassTech.
- **Pilot User Recruitment:** MassTech will **identify and recruit pilot users** from the stakeholder community, and the vendor will collaborate on the outreach.
- **Security Standard Alignment:** The DCC will adhere to the **data classification security standards** set by the MA CISO. Governance must also comply with applicable state and U.S. federal laws, and vendors are encouraged to align with standards such as the **NIST AI Risk Management Framework**.

3. Governance and Personnel Assumptions

- **Onshore Requirements:** Key project functions involving **governance, security, compliance, and stakeholder engagement** must be performed by **U.S.-based personnel**.
- **Policy Definition:** Defining governance policies and standards (e.g., ethical usage, bias detection, data tier delineation) will be a **collaborative effort** with MassTech and its established governance committee.
- **Start Date:** The anticipated project start date is **December 2, 2025**.

4. Out-of-Scope Items (Exclusions and Future Phase Deferrals)

The following activities, functionalities, and requirements are explicitly considered out of scope for the current phase of the project:

4.1 Commons Credits and Financial Systems

- **Real-Currency Transactions:** The Commons Credits system will operate as a **closed-loop incentive system** within the DCC, **not a real-currency mechanism**. Real-currency purchases are **out of scope initially**.
- **Credit Trading and Currency: Peer-to-peer credit trading** between users and **multi-currency support** (e.g., credits to USD exchange) are **not in scope**.
- **External Financial Integration:** Integration with **COMMBUYS** or performing **accounting and invoicing for cross-agency cost allocation** is **out of scope**.
- **Federal Grant Integration:** Integration with federal funding sources (e.g., NIH, NSF grants) for research usage tracking is **out of scope** for this phase.

4.2 Data Access, Storage, and Operations

- **Offshore Data Access: Remote access to Commonwealth Data from outside the continental United States** is strictly **prohibited** unless approved in advance in writing by the Commonwealth.
- **Vendor Data Storage:** The contractor is **not responsible for storing any data** in the contractor's computing environment.
- **24/7 Operational Support:** Ongoing **24/7 system support** is **not in scope** for this RFP; the vendor will provide support through the initial rollout phase.
- **Streaming Data: Real-time feeds and streaming datasets** (e.g., Kafka/Event Hub integration) are **out of scope** for the initial phase of work, planned for future releases (Year 2 or 3).

- **Offline/Low-Bandwidth Modes:** Support for **offline or low-bandwidth modes** for mobile users is **not in scope** for this phase.
- **Archival Strategy:** An archival strategy for logs **beyond three years** has not been defined.

4.3 AI and Fairness Capabilities

- **Real-Time Bias Monitoring: Real-time monitoring** of deployed models for fairness drift is **not required** for this phase; the architecture should include hooks for future capability.
- **Fairness Reporting for Public: Public-facing fairness dashboards** are **not required** for the initial phase; reporting is primarily for internal auditing and compliance.
- **Interactive Scenario Testing: Interactive scenario testing** (e.g., What-If Tool functionality) is **not in scope** for this phase.
- **Synthetic Data Production Grade:** Synthetic data pipelines are intended primarily for **R&D and sandbox environments**, not full production-grade usage.

4.4 Governance and Project Scope

- **External Site Access:** Access to **other MassTech sites and platforms** (e.g., Single Sign-On for sites outside the DCC) is **out of scope**.
- **Third-Party Certification:** A formal **third-party accessibility audit or certification** is **not mandated** as a baseline requirement in the RFP.
- **Multilingual UI at Launch:** While the platform must meet WCAG 2.1 AA accessibility standards and the architecture must enable future localization, **multi-language support** (English + others) is **not strictly required in Phase 1**.
- **Mandated Technologies:** The tools listed in the RFP (e.g., CKAN, Dataverse, SDV, Synthea, AIF360, Fairlearn) are **illustrative examples**, and **not mandatory components** of the technology stack. Vendors are free to propose alternatives.

XI. Past Performance & References

Case Study 1
Knowledge Fabric for Technology Disruption Framework

CONTACT PERSON	Chief Technology officer (CTO)
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

CLOUDSUFI successfully provided Data Engineering Services for Knowledge Fabric, supporting hedge fund's AI's product data consumption and ongoing data updates. Services, primarily in the US and London with remote support, established a robust foundation for research data, focusing on the Enterprise Knowledge Graph with Validation Agent.

Scope & Accomplishments

Data Corpus enrichment: Identified and validated data sources (BNEF, World Economic data, IRENA). Developed automated 1000+ pipelines and ETLs. Built the Stellar Knowledge Fabric using a Graph Database for an Enterprise Knowledge Graph.

Data Validation agent

Created, trained, and deployed a Validation Agent model to ensure data accuracy. Digitized all interconnected data assets within the Knowledge Fabric.

Business Benefits

- Enhanced Data Accuracy and Reliability: Validation Agent ensures high data quality.
- Accelerated Investment Decisions: Enterprise Knowledge Graph provides comprehensive market data for quicker, informed choices.
- Operational Efficiency: Automated processes streamline data handling, reducing research timelines by 90%.

Case Study 2
Patient 360 Platform

CONTACT PERSON	Chief Information officer (CIO)
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

Overview

CloudSufi, with GCP and DataCommons.org, built a **FAIR-compliant Patient 360 Platform** for a U.S. healthcare provider. The platform unified fragmented patient data (EMR, claims, labs, pharmacy, wearables, and social determinants) into a secure, AI-ready environment.

Challenge

Data silos, poor metadata, compliance risks (HIPAA), and limited AI capabilities prevented holistic patient insights and slowed research.

Solution

- **Platform:** Multi-tenant data lake and metadata catalog on GCP with DataCommons.org schema.
- **AI/ML:** Predictive risk scoring, outcomes forecasting, and synthetic data generation with Vertex AI.
- **Governance:** HIPAA-compliant security, audit trails, and fairness/explainability frameworks.

Impact

- 15% reduction in readmissions.
- 20% cut in admin overhead via automation.
- 90% decrease in reporting errors.
- FAIR-compliant catalog enabling faster research (minutes vs. weeks).

Case Study 3
Industry : Fashion and Jewelry

CONTACT PERSON DESIGNATION	Director of Data Engineering & AI
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

Challenge

- **Challenge 1:** Fragmented Data Landscape, data trapped across multiple, disconnected legacy ERP systems
- **Challenge 2:** Rigid, monolithic legacy ERP architecture was unable to support modern needs, limiting capabilities for real-time analytics and AI applications

Impact

- **Established Single Source of Truth:** Successfully implemented 1000+ data objects on BigQuery, creating a trusted, consistent foundation for all forward-looking business decisions and enterprise reporting
- **Accelerated Migration with AI:** Leveraging our in-house AI agents to automate complex transformation logic and mapping, reducing the time and manual effort traditionally required for large-scale data migration

Business Benefits

- **Empowered Strategic Decision-Making:** Shifted business reporting from retrospective analysis to forward-looking, scenario-based insights, ensuring leaders can make proactive decisions that directly impact revenue and operational efficiency

Case Study 4
Data Lake Modernisation

CONTACT PERSON	Senior Product Manager (Google Cloud)
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

Challenge

- **Challenge 1:** Customers had to develop their own custom solution to ingest data into BigQuery and also to publish the outbound data to external target systems. This challenge was addressed by standard managed connectors developed to cater to all the critical Target systems.
- **Challenge 2:** Templatized, intuitive, simplified UI based configuration driven connectors to connect to any source/sink to/from Google BigQuery

Impact

- **High-Performance Data Integration Connectors:** Developed SaaS connectors with 100% object coverage, industry-leading throughput, continuous enhancements based on enterprise feedback, and comprehensive maintenance/implementation guides.
- **Robust Data Validation Framework:** Created a specialized Test Framework for complex ETL data integrity, handling intricate data type mappings, and implementing CI/CD for automated validation.

Business Benefits

- **Seamless Third-Party Integrations:** Ensures high data quality and throughput for complex SaaS ETL (batch and real-time).
- **Guaranteed Data Integrity:** Our Validation Framework simplifies ETL maintenance and ensures data accuracy via periodic upgrades.
- **Enhanced Operational Efficiency:** Implementation guides significantly reduce customer overhead during large deployments.

XII. Bios of Key Personnel

#	NAME	ROLE TITLE	KEY RESPONSIBILITIES	REQUIRED SKILLS & TECHNOLOGIES
CLOUDSUFI				
1	Irfan Khan	Executive Sponsor	Provide executive oversight, ensure alignment with MassTech leadership, facilitate escalation and partnership with Google Cloud	Executive leadership, Strategic Partnerships, Stakeholder Engagement
2	Walid Hassan	Technical Program Manager	Lead program planning and execution, manage timelines and deliverables, run Agile ceremonies, provide risk and change management, ensure stakeholder communication.	Agile, Scrum/SAFe, Jira, Confluence, Risk Management, Cloud Program Delivery
3	Chiranjeev Dahiya	Business Analyst	Engage stakeholders to capture business and functional requirements, translate needs into user stories, validate against FAIR data principles, support testing documentation	Requirements Gathering, Agile, JIRA, Metadata Standards, FAIR Principles, Documentation
4	Vasudev Nayak	Cloud Solution Architect	Design and oversee end-to-end DCC architecture on Google Cloud; ensure scalability, security, and compliance; guide technical teams on integration patterns.	Google Cloud Architecture (BigQuery, Dataflow, IAM, Vertex AI), Security & Compliance, API Management
5	Amit Chakraborty	Integration Expert	Build and maintain secure data ingestion pipelines, integrate metadata catalog and synthetic data tools, ensure seamless interoperability with APIs and external datasets.	Workflow Orchestration (Dataflow, Airflow), Application Integration, CKAN/Dataverse, Synthetic Data Tools (SDV, Synthea, CARLA), API Management

Google Cloud				
7	Annette Ball	Account Executive - Google Public Sector	Serve as Google's executive liaison to MassTech; ensure access to the right Google resources; facilitate contract alignment, licensing, and strategic initiatives for public sector engagements.	Public Sector Account Management, Strategic Partnerships, Google Cloud Contracts & Licensing, Stakeholder Engagement
8	Mina Behnam	Senior Customer Engineer Google Cloud	Provide deep technical expertise on Google Cloud services; support solution architecture, security/compliance mapping, and performance optimization; ensure MassTech benefits from best practices and Google's innovation roadmap.	Google Cloud Architecture (BigQuery, Vertex AI, IAM, Dataflow), Security & Compliance (FedRAMP, HIPAA), Data & AI Best Practices, Cloud Cost Optimization

Proposals for

AI Hub Data Commons Collaborative

CONSULTING SERVICES

RFP No. 2026-JAII-01

Massachusetts Technology Collaborative

75 North Drive
Westborough, MA 01581-3340
http://www.masstech.org

Submitted by : CLOUDSUFI Inc.

07 OCT 2025

Table of Contents

I. Executive Summary & Partnership Vision **2**
IA. Why Google cloud ?... Google's Work in Data Commons 4
1. Data Integration and Standardization 4
2. Tools and Accessibility 4
3. AI Grounded in Facts 5
4. Leader in Artificial Intelligence 5
5. Commitment from the Extended Google Team 5
6. Google Executive Briefings 6
II. Profile & Qualification **7**
1. Expertise in: Data Commons / Cataloging Platforms (CKAN, Dataverse) 9
2. Expertise in: Metadata Management and FAIR Standards 9
3. Expertise in: Synthetic Data (tools like SDV, Synthea, CARLA) 10
4. Expertise in: AI Fairness (AIF360, Fairlearn) 10
5. Expertise in: Security/Governance Frameworks (RBAC, consent management, NIST standards) 11
6. Google Expertise: Unique Characteristics of Google 12
More than just world-class infrastructure 12
III. Proposed Solution Architecture: The Integrated Data Commons **15**
IV. Technical Capabilities: Core Infrastructure & Cataloging **19**
1. Presentation and Access Layer 20
2. Control Plane & Governance 21
3. Data Core & Curation Engine 22
4. Central Storage & MLOps Platform (Zone A) 22
5. Zone B: Hosted/On-Prem - High-Performance Compute & AI 23
V. Enabling Financial Governance: The Commons Credits System **26**
VI. Differentiating AI Capabilities **29**
VI.A. Synthetic Data Generation (RFP 2.4) 29
VI.B. Responsible AI and Bias Framework (RFP 2.5) 30
VI.C Deeply Embedded in Google Cloud's Innovation Efforts 31
VII. Security, Governance, and Compliance Framework **32**
VIII. Project Management, Delivery, and Ecosystem Growth **34**
Detailed Project Plan (40 Weeks) 36
Project Governance & Change Management 43
RFP Requirements Compliance Checklist (Phase Integration) 44
IX. RFP Requirements Compliance Checklist **46**
X. Assumptions **49**
XI. Past Performance & References **52**
Knowledge Fabric for Technology Disruption Framework 52
Patient 360 Platform 53
Industry : Fashion and Jewelry 54
XII. Bios of Key Personnel **56**

I. Executive Summary & Partnership Vision

CLOUDSUFI, leveraging the power and proven architecture of **Google's Data Commons** and partnering with **Google Cloud Platform (GCP)**, proposes a strategic partnership to establish a world-class **Data Commons Collaborative (DCC)** for the Massachusetts AI Hub. This solution is designed for **rapid implementation, superior security, accessibility, and unprecedented value**, directly addressing the RFP's core evaluation criteria.

The foundation of our proposal is **Google Cloud's operational Data Commons infrastructure**—the only cloud-native knowledge graph of its kind among major providers. Rather than building a bespoke platform from scratch, we will deliver a **secure, customized instance** of this already proven technology. This approach accelerates the timeline, aligns with Massachusetts' vision for a statewide Data Commons, and reduces the risks that typically arise in large-scale data integration and harmonization efforts.

To ensure precision, transparency, and optimal value for the Commonwealth, **CLOUDSUFI and Google Cloud are jointly proposing an initial investment phase - two weeks of on-site workshops at zero cost to MassTech.** During this phase, our combined teams will work closely with stakeholders to document existing systems, validate data flows, and refine infrastructure sizing and cost models based on real-world conditions. By absorbing this upfront effort, we **de-risk implementation** for the Commonwealth and ensure that subsequent cost estimates reflect verified requirements rather than assumptions—demonstrating our **commitment to partnership, accountability, and long-term success.**

The proposed platform extends far beyond data aggregation. We deliver a **cohesive, serverless, and secure ecosystem** built on best-in-class, natively integrated Google services:

- **BigQuery** for petabyte-scale analytics
- **Vertex AI with Gemini models** for advanced modeling and synthetic data
- **Apigee** for secure API management with **API keys, rate-limits, and quotas** to ensure fair access
- **Dataplex** for governance and metadata
- **Looker** for transparency, reporting, and dashboards

This integrated stack provides lower total cost of ownership, higher scalability, and less operational risk than fragmented, open-source alternatives (e.g., CKAN, Dataverse).

The CLOUDSUFI–Google partnership also brings **unique credibility**. We are actively co-developing **governance capabilities on Dataplex** and hold the **three-year Google Data Commons Data Loading SOW**, underscoring our operational role in the same knowledge graph Massachusetts will adopt. This

ensures that the Commonwealth is not betting on untested tools but aligning with the **roadmap powering Google's global data innovation**.

Critically, the DCC will transform into a **living knowledge graph** with **semantic search and natural language (NLP) interfaces available in Phase 1**, allowing researchers, policymakers, and innovators to pose questions in everyday language rather than navigate schemas. Coupled with Google's **Responsible AI (RAI) toolkit** for fairness, explainability, and bias mitigation, and our **Commons Credits system** for transparent, equitable allocation of resources, the platform ensures broad participation and long-term sustainability.

Together, **Google Cloud and CloudSufi** will deliver not just a system that meets the RFP's requirements, but a **strategic asset that positions Massachusetts as a national leader in responsible, AI-driven innovation and economic development.**

Compliance Highlights at Launch: The DCC will provide **semantic search and natural-language access in Phase 1**, operate across **Dev/Test, QA, and Production** environments, and expose a secure **API gateway (Apigee) with API keys, rate-limits, quotas, and monitoring**. The public web experience will meet **WCAG 2.1 Level AA** with support for screen readers, keyboard navigation, and captioning. Audit logs will be retained for **at least three years** with options for archival expansion as policies are finalized.

IA. Why Google cloud ?... Google's Work in Data Commons

Google created and maintains Data Commons as an open-source platform whose mission is to organize the world's publicly available statistical data and make it universally accessible and useful.
Google's work for Data Commons is to act as a central data integrator, transformer, and distribution point, making complex, siloed public information easy to access and leverage, especially with AI.
Outlined below are three main areas summarizing the work that Google does for Data Commons.

1. Data Integration and Standardization

The most fundamental work Google does is taking vast amounts of fragmented public data and uniting it into a cohesive resource. A few key activities include:

- Creates a Unified Knowledge Graph: We collect data from over 200 trusted sources, including the U.S. Census Bureau, the World Bank, the United Nations, and various government and non-governmental organizations (NGOs).
- Data Normalizing and Cleaning: We perform the tedious, costly work of normalizing, cleaning, and reconciling disparate data. This means aligning different schemas, formats, and references so that data about the same entity from different sources can be accessed instantly through a single query.
- Non-Ownership: Lastly, we do not collect or own any of the data itself - we simply aggregate, process, and present publicly available information and provide the link back to the original source.

2. Tools and Accessibility

We develop multiple user-friendly tools to make the data accessible to everyone. We do this by:

- APIs and Libraries: We provide programmatic access via REST APIs and a Python client library and this enables data scientists and developers to integrate Data Commons into their custom projects.
- Web-Based Explorers: For non-coders this provides tools on the website (e.g. Map Explorer, Statistical Variable Explorer, and Scatter Plot Explorer) to analyze, visualize, and compare data without writing code.
- Google Search Integration: A subset of Data Commons' data is directly integrated into Google Search. When a user asks a statistical question (e.g., "What is the population of New Zealand?"), the factual answer, often presented as a knowledge graph or chart, is retrieved directly from Data Commons.

3. AI Grounded in Facts

In recent years, we have focused on using Data Commons as a trusted "data backbone" for our AI models.

- Mitigating AI Inconsistencies: Data Commons is used to ground Large Language Models (LLMs) in real-world, verifiable statistical data. This technique helps reduce the rate of "hallucinations" (inaccurate or invented information) by providing the model with sourced facts.
- Natural Language Querying: We developed the Model Context Protocol (MCP) Server to allow AI agents to consume Data Commons natively. This means developers can query the vast dataset using simple, natural language questions, without having to master complex APIs.
- AI Model Training: Models like our open DataGemma are specifically built to leverage the Data Commons knowledge graph to enhance their factuality and reasoning. This serves as an explicit bridge between LLMs and real-world data.

4. Leader in Artificial Intelligence

Google holds a prominent leadership position in Artificial Intelligence (AI), driven by decades of foundational research and the creation of cutting-edge AI models. Our leadership is anchored by Google DeepMind, which is responsible for landmark breakthroughs like AlphaGo and the protein-folding system AlphaFold, alongside the development of its most powerful and multimodal large language model, Gemini. Google has integrated our AI at every level, from its specialized Tensor Processing Units (TPUs) for accelerated model training, to core product features like AI Overviews in Search, and the Vertex AI platform, that allows enterprises to build and deploy their own AI solutions.

5. Commitment from the Extended Google Team

We are committed to bring ‘the best of Google’ to support the Massachusetts Technology Collaborative and the Data Commons project. Outlined below in Figure 1 are some of the key resources that Google will bring to the partnership.

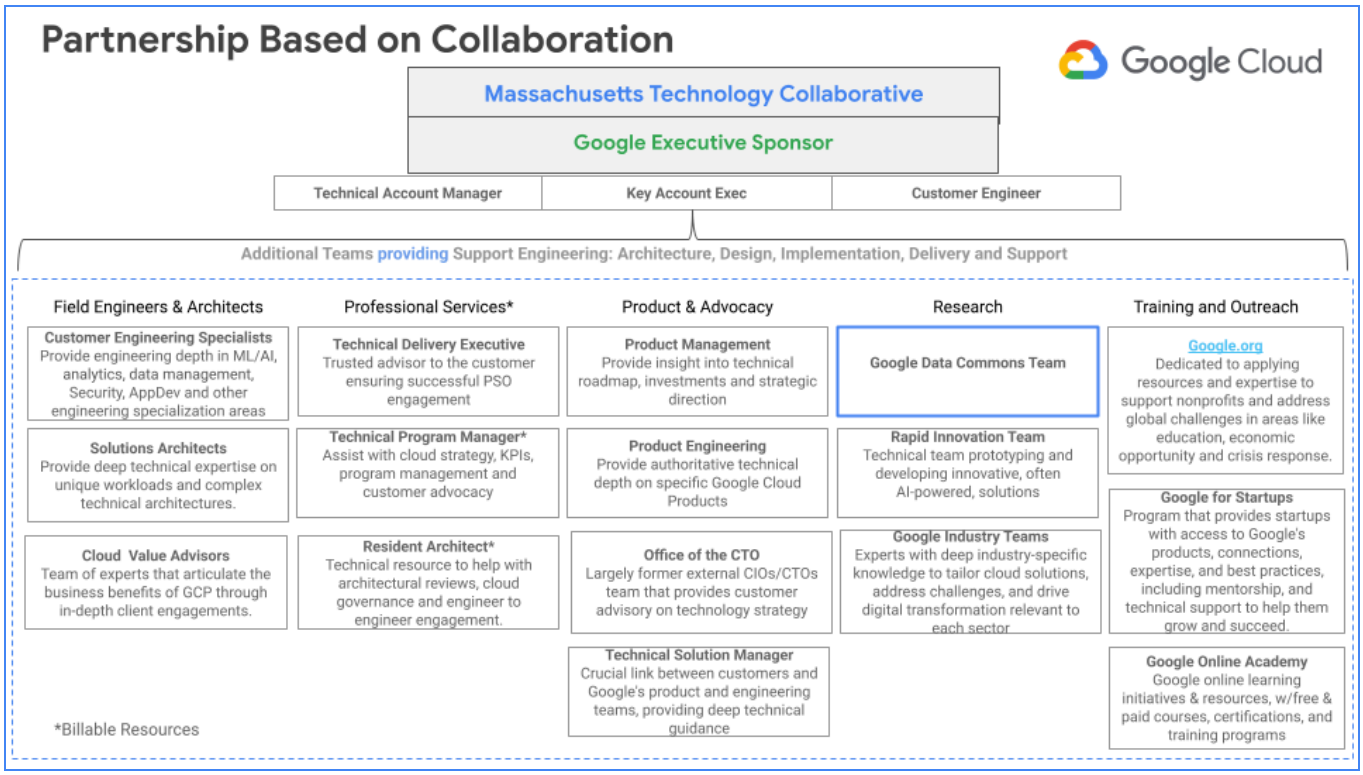

Partnership Based on Collaboration
Google Cloud
Massachusetts Technology Collaborative
Google Executive Sponsor
Technical Account Manager
Key Account Exec
Customer Engineer
Additional Teams providing Support Engineering: Architecture, Design, Implementation, Delivery and Support
Field Engineers & Architects
Professional Services*
Product & Advocacy
Research
Training and Outreach
Customer Engineering Specialists
Provide engineering depth in ML/AI, analytics, data management, Security, AppDev and other engineering specialization areas
Technical Delivery Executive
Trusted advisor to the customer ensuring successful PSO engagement
Product Management
Provide insight into technical roadmap, investments and strategic direction
Google Data Commons Team
Google.org
Dedicated to applying resources and expertise to support nonprofits and address global challenges in areas like education, economic opportunity and crisis response.
Solutions Architects
Provide deep technical expertise on unique workloads and complex technical architectures.
Technical Program Manager*
Assist with cloud strategy, KPIs, program management and customer advocacy
Product Engineering
Provide authoritative technical depth on specific Google Cloud Products
Rapid Innovation Team
Technical team prototyping and developing innovative, often AI-powered, solutions
Google for Startups
Program that provides startups with access to Google's products, connections, expertise, and best practices, including mentorship, and technical support to help them grow and succeed.
Cloud Value Advisors
Team of experts that articulate the business benefits of GCP through in-depth client engagements.
Resident Architect*
Technical resource to help with architectural reviews, cloud governance and engineer to engineer engagement.
Office of the CTO
Largely former external CIOs/CTOs team that provides customer advisory on technology strategy
Google Industry Teams
Experts with deep industry-specific knowledge to tailor cloud solutions, address challenges, and drive digital transformation relevant to each sector
Technical Solution Manager
Crucial link between customers and Google's product and engineering teams, providing deep technical guidance
Google Online Academy
Google online learning initiatives & resources, w/free & paid courses, certifications, and training programs
*Billable Resources

Figure 1. Google Team Supporting the Data Commons Project in addition to CLOUDSUFI Inc.

6. Google Executive Briefings

As part of the partnership, Google welcomes the opportunity to host an annual executive briefing with Massachusetts Technology Collaborative staff, CLOUDSUFI, and the extended Google team. These briefings can serve as a platform to bring us together to discuss technology developments that can be leveraged to enhance the data commons project, product roadmap enhancements, and other key topics that are important to the parties.

II. Profile & Qualification

CLOUDSUFI Profile	
SECTION	DETAILS
Company Name	Cloudsufi Inc.
Headquarters	San Jose, California, United States
Founded	2019
Legal Status	LLC
Website	www.cloudsufi.com
Vision	Eliminate the gap between “Human Intuition” and “Data-driven Decisions
Core Competencies	- Product Engineering - Platform, Mobile , Web - AI , ML & Data Engineering - Advanced Analytics - Data Commons & Metadata Catalogs - AI Fairness & Responsible AI Frameworks - Cloud-Native Architecture & Governance - Security & Compliance (RBAC, MFA, HIPAA, SOC2, FedRAMP experience)
Strategic Partnerships	- Google Cloud (Preferred Technology Partner) - Workato, Oracle,SAP
Certifications	- Google Cloud Premier Partner - Google Cloud Certified Engineers & Architects - ISO/SOC2 readiness and security compliance experience
Relevant Experience	- Delivered large-scale data platforms - Implemented AI-driven synthetic data solutions - Built governance frameworks with FAIR-compliant data access

CLOUDSUFI Profile	
SECTION	DETAILS
Leadership & Team	- Program Managers with experience in multi-stakeholder AI/Cloud programs - Google Cloud–certified Architects & Engineers - Specialists in governance, fairness, and compliance
Differentiators	- Exclusive alignment with Google Cloud AI & Data services - Strong expertise in ethical AI and bias mitigation - Proven phased delivery methodology - Cost-efficient, scalable cloud-native designs
Corporate Responsibility	- Commitment to responsible AI and inclusion - Supports sustainable cloud operations - Focus on equitable access to AI innovation in Massachusetts and beyond

Name of Technology Partner & Sponsor – **Google Cloud**	
Customer	Massachusetts Technology Collaborative
RFP Title	AI Hub Data Commons Collaborative Consulting Services
RFQ No.	2026-JAII-01
Submitted by	Annette Ball
Phone No.	(518) 429-4896
Email	annetteball@google.com
Capture Manager	Mark Slafka | mslafka@google.com | (703) 483-1416

CLOUDSUFI with Google Cloud approaches the DCC not merely as a technology implementation but as the creation of a sophisticated, extensible platform designed to transform data silos into a knowledge-sharing ecosystem. Our proposal focuses on leveraging best-in-class technologies, whether open-source or proprietary, to deliver a solution that meets the Commonwealth's goals for scalability, security, and ethical AI development.

1. Expertise in: Data Commons / Cataloging Platforms (CKAN, Dataverse)

The RFP requires the implementation of **metadata management, cataloging, and discovery tools**, explicitly suggesting that the **CKAN** and **Dataverse** platforms be considered.

- **RFP Context:** CLOUDSUFI understands that the tools listed (including CKAN, Dataverse, DataHub, and Amundsen) are provided as **illustrative references** and are **not mandated solutions**. Vendors are encouraged to propose platforms that are the best fit based on criteria like scalability and interoperability.
- **CLOUDSUFI Approach:** Our strategy focuses on designing a superior, deeply integrated metadata solution that provides the required capability for dataset discovery. We will propose an approach that supports both **auto-harvesting from external sources** (e.g., existing CKAN/Dataverse catalogs) and rigorous **manual curation** workflows to ensure quality and governance. The resulting architecture will support a dataset directory that is **searchable and filterable** by key criteria such as data type, domain, source, and file format.

2. Expertise in: Metadata Management and FAIR Standards

A core requirement of the DCC is ensuring **metadata completeness** and rigorous adherence to **FAIR (Findable, Accessible, Interoperable, Reusable) data principles**.

- **FAIR Implementation:** CLOUDSUFI recognizes that compliance with FAIR principles, supported by widely used frameworks such as **DCAT** and **Schema.org**, is essential to ensure that datasets are discoverable and reusable.
- **Metadata Scope:** Our solution will manage the three required categories of metadata: **Descriptive Metadata** (Title, license), **Technical Metadata** (File format, schema), and **Provenance** (who processed the data).
- **Lineage and Versioning:** We will ensure that the initial launch includes, at a minimum, **dataset-level lineage tracking** to capture the origin and transformations of the entire dataset. Our roadmap will include the capability to incorporate more granular **column-level lineage** in future phases to meet evolving transparency needs. Metadata records themselves will also be **versioned** to maintain a history of schema and description changes.

3. Expertise in: Synthetic Data (tools like SDV, Synthea, CARLA)

The RFP mandates the integration of a **Synthetic Data Capability** to support model training and testing in privacy-preserving and scalable ways.

- **Tooling:** CLOUDSUFI is highly experienced with the open-source tools cited—**SDV** (for tabular data), **Synthea** (for health records), and **CARLA** (for simulations)—but recognizes these are provided only as **illustrative examples**. We are prepared to recommend the most appropriate tools and frameworks (custom, existing, or generative AI-based) to meet the DCC's needs.
- **Scope and Priority:** The synthetic generation capability must support **multimodal** data (image/video/audio) and allow for **configurable privacy-utility tradeoffs**. Priority domains for synthetic data pilots include **life sciences, robotics, education, and healthcare**.
- **Governance:** The platform must include **audit trails** linking synthetic datasets back to the original real datasets for validation and compliance purposes. Furthermore, the pipelines must align with applicable standards such as **HIPAA, GDPR, and relevant Massachusetts state privacy laws**. MassTech is open to full-scale synthetic derivative datasets for production AI applications, not just demonstration sets.

4. Expertise in: AI Fairness (AIF360, Fairlearn)

Ensuring ethical and equitable AI development requires a robust fairness and bias framework, which must include implementation of **bias detection, auditing, and mitigation tools**.

- **Tooling:** Like the synthetic data tools, **AIF360** and **Fairlearn** are listed as illustrative examples of the required capability. CLOUDSUFI's expertise encompasses not just these specific libraries but a comprehensive toolkit that supports **dataset auditing, bias mitigation, model explainability, and drift detection**.
- **Scope:** The platform must support fairness checks for both **models and datasets**. The Massachusetts AI Hub is concerned with a broad range of biases (data, algorithmic, societal) and requires evaluation across protected attributes such as **race, gender, age, and disability**.
- **Monitoring and Reporting:** The solution must support the **continuous monitoring of fairness drift in deployed models**. For this phase of work, fairness reporting is intended primarily for **internal auditing and compliance**, though public transparency features will be supported. Governance should align with standards such as **NIST AI Risk Management Framework**.

5. Expertise in: Security/Governance Frameworks (RBAC, consent management, NIST standards)

Security is the foundation of the DCC, demanding adherence to state and federal requirements, including robust access control and auditability.

- **Access Control (RBAC & MFA):** We will design and implement the required **Role-Based Access Control (RBAC)** to manage access across the required dataset tiers (Public, Restricted, Internal). All sensitive data access will be protected by **Multi-Factor Authentication (MFA)**.
- **Data Residency and Security:** The architecture will ensure that all data is **encrypted at rest and in transit**. All data processing must occur in **U.S. regions with an East-coast preference**, and remote access to Commonwealth Data from outside the continental U.S. is strictly **prohibited**. The DCC must adhere to **MA CISO data classification security standards**.
- **Consent Management:** CLOUDSUFI will implement the required dynamic and verifiable **consent management system in Phase 1**. This system will ensure that consent, primarily for restricted datasets, is **tokenized, revocable, and time-bound**.
- **Auditing and Compliance:** The system is required to **log all data access, downloads, and synthetic generation events**. These audit logs must be retained for at least **three years** for compliance review. We will propose security architectures that align with widely recognized best practices and can adapt to frameworks like the **NIST AI Risk Management Framework**.

6. Google Expertise: Unique Characteristics of Google

More than just world-class infrastructure

- **Infrastructure as Scale**

  The Massachusetts Technology Collaborative will leverage Google's unparalleled backbone network. This network, which moves at "Google speed" — near-zero latency — for billions of users worldwide is now available to governments and enterprises.

- **Advanced Analytics**

  We bring more than just infrastructure, but distinct capabilities in leveraging unparalleled leadership in advanced analytics and intelligent automation, built upon foundational technologies that redefine how state agencies operate.

  We offer world-class capability to unify and extract deep, actionable insights from vast and diverse government datasets, transforming raw information into actionable intelligence that fuels smarter policy and more effective program delivery. Furthermore, our pioneering approach to highly scalable and resilient application environments provides a robust foundation for developing agile and future-proof digital services, ensuring continuous delivery even as demands evolve.

- **Google Cloud Security**

  As a cloud innovator, Google understands security in the cloud. Our cloud services are designed to deliver better security than many on-premises approaches. We make security a priority in our operations—operations that serve users across the world. Security drives our organizational structure, culture, training priorities, and hiring processes. It shapes the design of our data centers and the technology that they house. It's central to our everyday operations and disaster planning, including how we address threats. It's prioritized in the way we handle customer data, our account controls, our compliance audits, and our certifications.

- **Commitment to Sustainability**

  Google Cloud is committed to leading the technology industry in sustainability, not as a separate initiative, but as a core tenet of our business. Our strategy is built on a foundation of ambitious goals, operational excellence, and a commitment to empowering our customers to achieve their own sustainability objectives. We believe that a sustainable cloud is a better cloud—more efficient, more resilient, and more responsible.

- **Proud to call Cambridge Home**

  Our Cambridge office has employees working across more than 30 teams, including our Global Business Organization (sales) and many different engineering and product areas. Some of the teams in Cambridge include Google Cloud, Flights & Hotels, Search & Assistant, networking infrastructure, Android, YouTube, and Chrome. We also have employees working on research in areas of machine learning and artificial intelligence.

Other Google Incentives for this program
Google Value-Added Offerings

- Fixed Price with Google's Public Sector Subscription Agreement

  To address the challenge that public entities face of having pre-defined budgets that cannot be exceeded for a project, we are pleased to offer the Public Sector Subscription Agreement (PSSA). This unique offering from Google was created to address these challenges by enabling our state and local government and education customers to purchase Google Cloud services at a fixed monthly price for a defined workload, regardless of actual consumption. This allows public sector customers, such as universities and government agencies, to achieve predictable costs, eliminate queues for resources, and easily scale computing power for large projects.

  The PSSA provides access to Google Cloud's full suite of services, including computing, storage, networking, security, data analytics, and advanced tools in Artificial Intelligence (AI) and Machine Learning (ML).

- Google Public Sector - Rapid Innovation Team (RIT)

  This team is made of industry executives and technologists and they focus on identifying challenges that government customers face in their job and prototype solutions to address these challenges. Their mission is to bring the 'Google Culture of Innovation' to our customers through proving mission challenges are possible and packaging them to scale 10x. We define "Rapid" as delivery of a high-fidelity prototype within 6-8 weeks.

  Google would welcome the opportunity to collaborate with the Massachusetts Technology Collaborative on challenges they are facing and offer Google's point of view for how we might be able to assist.

- No Cost Training for Massachusetts Technology Collaborative

  Google provides comprehensive resources for career advancement through its Google Cloud Skills Boost and Grow with Google Career Certificates platforms.

Google Cloud Skills Boost enables individuals and teams to deepen their expertise in Google Cloud technologies, offering specialized training, skill badges, and industry-recognized certifications for roles like cloud architects, data engineers, and security specialists.

In parallel, Grow with Google Career Certificates delivers accessible, practical training in high-demand fields such as IT support, data analytics, project management, and UX design, designed for learners with no prior experience or traditional degrees. Both platforms equip users with the skills needed to thrive in job roles relevant to government agencies and educational institutions, fostering professional growth and career success.

- Explore Cloud Skills Boost
- Earn a Grow with Google certificate

III. Proposed Solution Architecture: The Integrated Data Commons

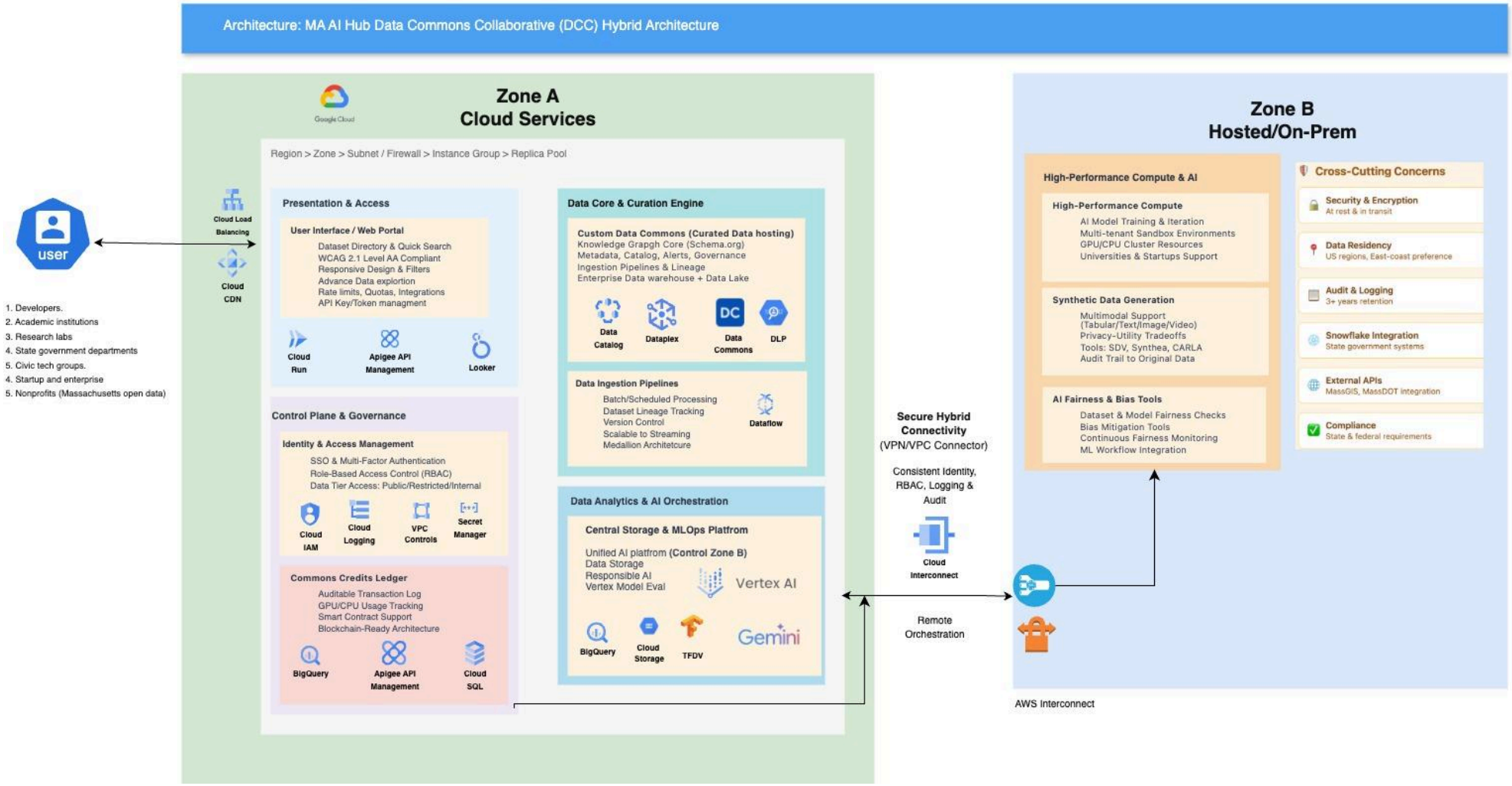
The CLOUDSUFI solution transforms the DCC from a static data repository into a **dynamic, interconnected living knowledge graph** for the Commonwealth. This technical architecture is based on the **proven, documented Data Commons reference architecture**. In addition to traditional catalog and query tools, the platform incorporates **semantic search and natural language (NLP) interfaces** that allow users to interact with the Data Commons using everyday language. Rather than requiring knowledge of table names or schemas, researchers, policymakers, and innovators can simply pose questions in natural language. Behind the scenes, the system applies **language models, embeddings, and ontology mapping** to translate those questions into structured queries and align terminology across datasets. These capabilities ensure the DCC is not only a secure and scalable repository, but also an **intuitive and inclusive discovery environment**, lowering technical barriers and accelerating insight generation across the Commonwealth's research and innovation ecosystem.

1. **Core Technology (Knowledge Graph):** The foundation is the Data Commons knowledge graph, which employs a standardized, open schema derived from Schema.org. Data is structured as semantic triples (entities, properties, relationships), allowing disparate datasets (e.g., census, economic, and health data) to be mapped to a single, canonical node (e.g., "Suffolk County"). This architecture facilitates immediate **cross-domain analysis** and eliminates time-consuming "data wrangling".

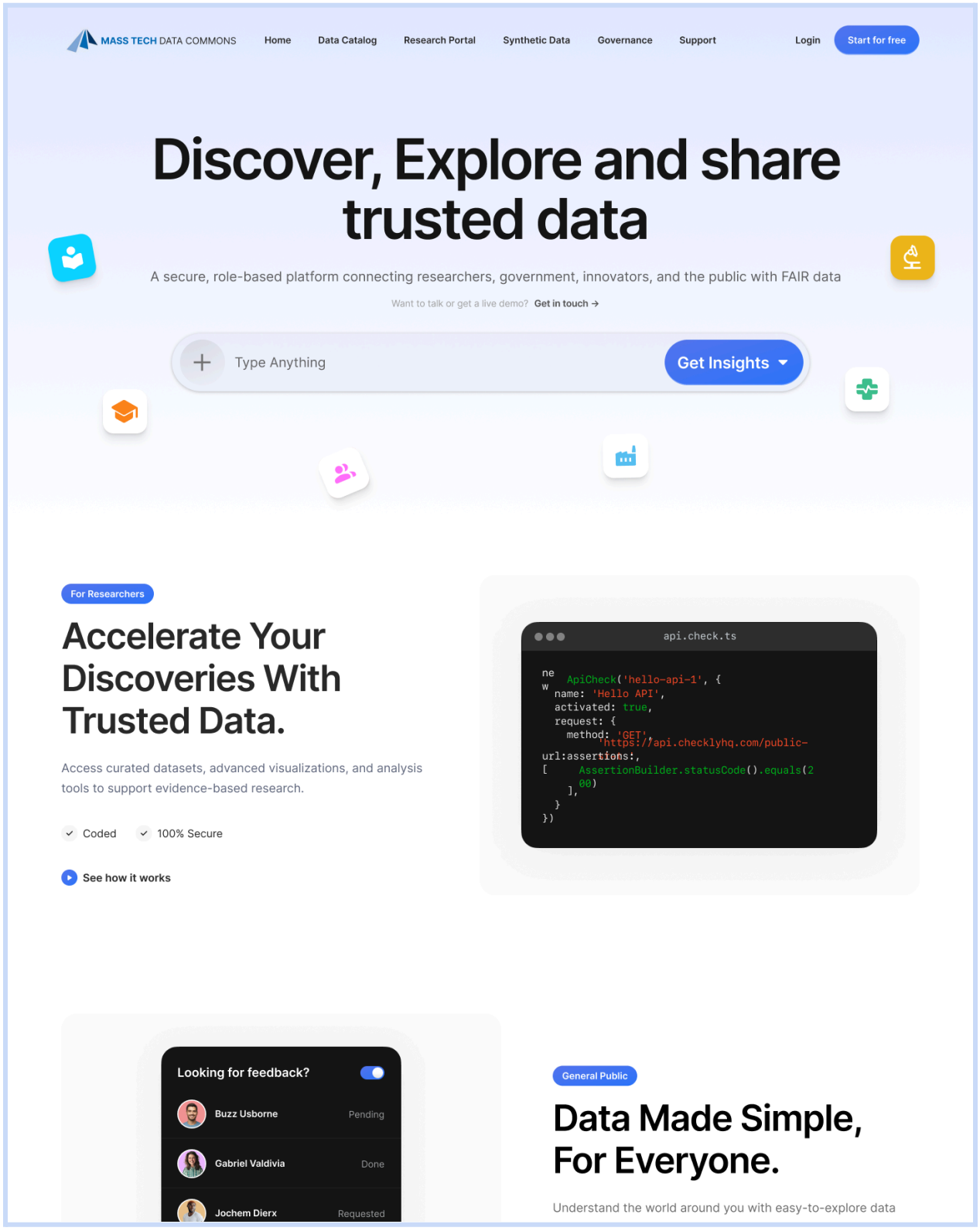
2. **FAIR Compliance:** The knowledge graph design is the core implementation of **FAIR principles** (Findable, Accessible, Interoperable, Reusable). The unified schema makes data inherently Interoperable, and its APIs ensure high Accessibility.

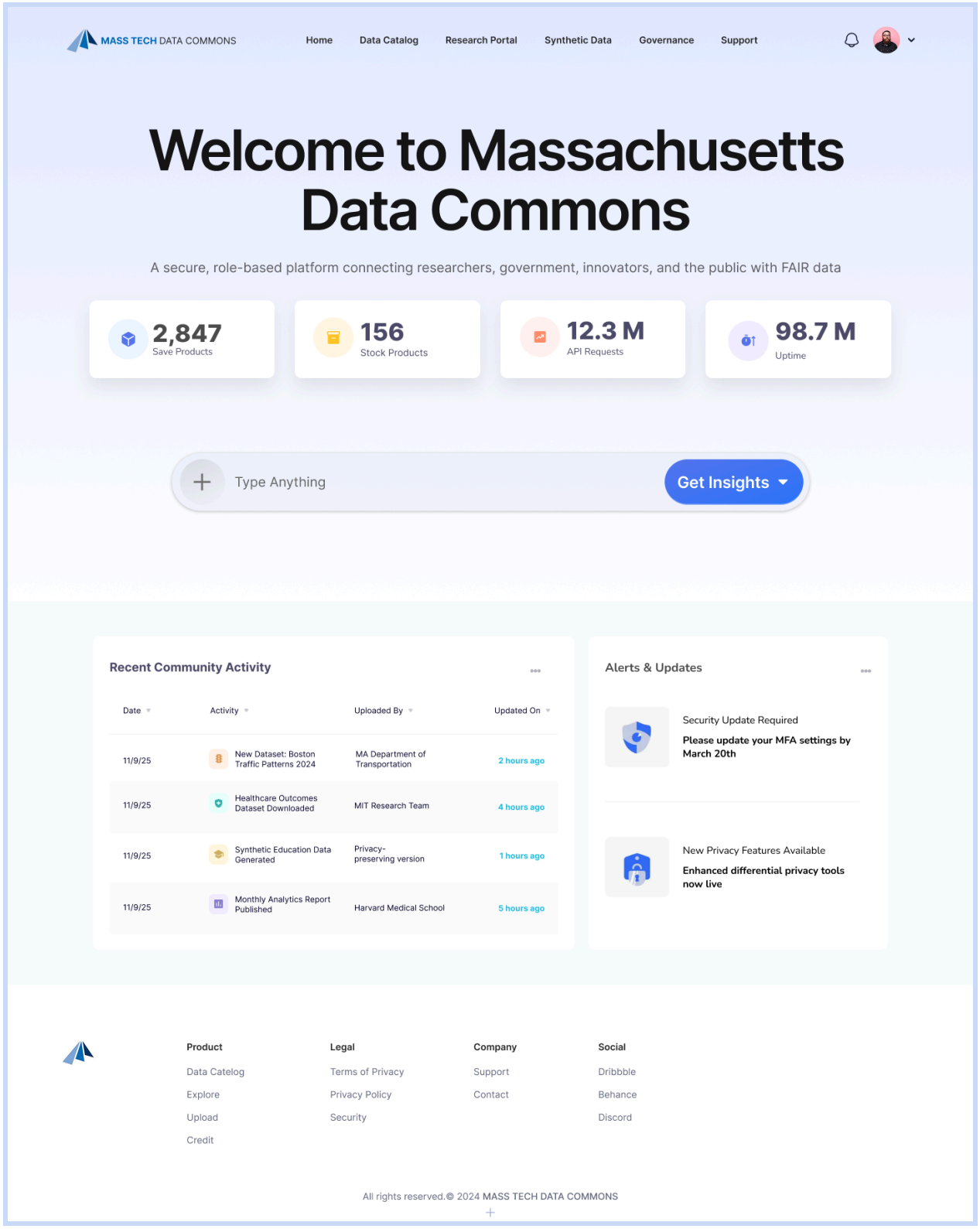
3. **Hybrid and Scalable Deployment:** The system is designed for a **hybrid deployment model**. The architecture will leverage **MGHPCC/AICR** for hosted data and compute, paired with **Cloud-based services** (GCP) for cataloging, APIs, and web access, ensuring consistent identity, RBAC, logging, and auditing across both domains. All data processing and storage will be in **U.S. regions** with an East-coast preference. The architecture is modular, API-first, and standards-aligned to enable seamless future expansion (e.g., multi-cloud federation, quantum computing integration).

High Level Architecture below


Architecture: MA AI Hub Data Commons Collaborative (DCC) Hybrid Architecture
Zone A
Cloud Services
Region > Zone > Subnet / Firewall > Instance Group > Replica Pool
user
1. Developers.
2. Academic institutions
3. Research labs
4. State government departments
5. Civic tech groups.
4. Startup and enterprise
5. Nonprofits (Massachusetts open data)
Cloud Load Balancing
Cloud CDN
Presentation & Access
User Interface / Web Portal
Dataset Directory & Quick Search
WCAG 2.1 Level AA Compliant
Responsive Design & Filters
Advance Data explortion
Rate limits, Quotas, Integrations
API Key/Token managment
Cloud Run
Apigee API Management
Looker
Control Plane & Governance
Identity & Access Management
SSO & Multi-Factor Authentication
Role-Based Access Control (RBAC)
Data Tier Access: Public/Restricted/Internal
Cloud IAM
Cloud Logging
VPC Controls
Secret Manager
Commons Credits Ledger
Auditable Transaction Log
GPU/CPU Usage Tracking
Smart Contract Support
Blockchain-Ready Architecture
BigQuery
Apigee API Management
Cloud SQL
Data Core & Curation Engine
Custom Data Commons (Curated Data hosting)
Knowledge Grapgh Core (Schema.org)
Metadata, Catalog, Alerts, Governance
Ingestion Pipelines & Lineage
Enterprise Data warehouse + Data Lake
Data Catalog
Dataplex
Data Commons
DLP
Data Ingestion Pipelines
Batch/Scheduled Processing
Dataset Lineage Tracking
Version Control
Scalable to Streaming
Medallion Architetcure
Dataflow
Data Analytics & AI Orchestration
Central Storage & MLOps Platfrom
Unified AI platfrom (Control Zone B)
Data Storage
Responsible AI
Vertex Model Eval
Vertex AI
BigQuery
Cloud Storage
TFDV
Gemini
Secure Hybrid Connectivity
(VPN/VPC Connector)
Consistent Identity, RBAC, Logging & Audit
Cloud Interconnect
Remote Orchestration
Zone B
Hosted/On-Prem
High-Performance Compute & AI
High-Performance Compute
AI Model Training & Iteration
Multi-tenant Sandbox Environments
GPU/CPU Cluster Resources
Universities & Startups Support
Synthetic Data Generation
Multimodal Support (Tabular/Text/Image/Video)
Privacy-Utility Tradeoffs
Tools: SDV, Synthea, CARLA
Audit Trail to Original Data
AI Fairness & Bias Tools
Dataset & Model Fairness Checks
Bias Mitigation Tools
Continuous Fairness Monitoring
ML Workflow Integration
Cross-Cutting Concerns
Security & Encryption
At rest & in transit
Data Residency
US regions, East-coast preference
Audit & Logging
3+ years retention
Snowflake Integration
State government systems
External APIs
MassGIS, MassDOT integration
Compliance
State & federal requirements
AWS Interconnect

Sample Prototype of **Data Commons landing page for MASSTech**




MASS TECH DATA COMMONS
Home
Data Catalog
Research Portal
Synthetic Data
Governance
Support
Welcome to Massachusetts Data Commons
A secure, role-based platform connecting researchers, government, innovators, and the public with FAIR data
2,847
Save Products
156
Stock Products
12.3 M
API Requests
98.7 M
Uptime
Type Anything
Get Insights
Recent Community Activity
Date
Activity
Uploaded By
Updated On
11/9/25
New Dataset: Boston Traffic Patterns 2024
MA Department of Transportation
2 hours ago
11/9/25
Healthcare Outcomes Dataset Downloaded
MIT Research Team
4 hours ago
11/9/25
Synthetic Education Data Generated
Privacy-preserving version
1 hours ago
11/9/25
Monthly Analytics Report Published
Harvard Medical School
5 hours ago
Alerts & Updates
Security Update Required
Please update your MFA settings by March 20th
New Privacy Features Available
Enhanced differential privacy tools now live
Product
Data Catelog
Explore
Upload
Credit
Legal
Terms of Privacy
Privacy Policy
Security
Company
Support
Contact
Social
Dribbble
Behance
Discord
All rights reserved.© 2024 MASS TECH DATA COMMONS

IV. Technical Capabilities: Core Infrastructure & Cataloging

The core infrastructure is fully managed and serverless, optimizing performance and cost.

COMPONENT	TECHNOLOGY	RATIONALE AND CAPABILITY DETAIL
Data Lake/Staging	Google Cloud Storage (GCS)	Highly durable, cost-effective object storage for raw datasets.
Analytical Engine	Google BigQuery	Fully managed, petabyte-scale data warehouse with a serverless architecture, separating storage and compute for performance and cost control.
Backend Compute/API	Cloud Run	Fully managed container platform for Data Commons application logic (APIs, web front-end). Automatically scales down (even to zero) based on traffic, ensuring cost-effectiveness.
Transactional DB	Cloud SQL	Managed relational database service for application data and transactional integrity.
Data Cataloging	Google Cloud Data Catalog	**Superior Solution:** Replaces standalone catalogs (CKAN/Dataverse) with deep cloud integration. Provides **automated metadata ingestion** from BigQuery and GCS. Uses **schematized tag templates** for structured business metadata and integrates seamlessly with **Cloud DLP** for automated identification and tagging of PII/sensitive data.
Discovery/ Analytics	Looker with Gemini	Looker's semantic modeling layer (LookML) ensures a **single source of truth** for metrics. **Gemini in Looker** enables **AI-powered analytics** via natural language querying, accelerating insight generation for non-technical policy

		analysts and researchers. **AI-driven discovery** (e.g., semantic search) is supported.
Lineage/ Versioning	Data Commons Import Tools	Provenance tagging is included. At launch, **dataset-level lineage tracking** is provided, with architecture designed for eventual **column-level lineage** expansion.
Google Cloud Dataplex	Centralize data governance, improve data quality	an intelligent data fabric that provides a unified, serverless, and automated approach to manage and govern data across an organization's distributed data lakes, data warehouses, and data marts. It helps data scientists and analysts to find and use high-quality data more easily

The architecture is designed as a **secure, hybrid, and integrated platform** that separates core cloud services (Zone A) from high-performance, potentially sensitive workloads (Zone B Hosted/On-Prem). This structure ensures governance, scalability, and adherence to the strict security and ethical requirements of the RFP.

The architecture is segmented into five primary functional layers, linked by secure connectivity and governed by a central control plane.

1. Presentation and Access Layer

This layer handles all user interaction, ensuring ease of use and compliance with accessibility standards.

- **User Interface:** A central **User Interface / Web Portal** is provided for interaction. It must be **WCAG 2.1 Level AA Compliant** and offer **Responsive Design**.
- **Discovery:** This portal provides a **Dataset Directory & Quick Search** capability, along with **Advanced Data Exploration** features, including semantic/NLP searches.
- **API Management:** Access to data and services is governed by **Apigee API Management**, which provides a secure and scalable gateway for all API interactions. Apigee manages **API keys/tokens**, ensuring only authorized access, and enforces **rate limits and quotas** to prevent abuse and maintain fair resource allocation across tenants. These policies can be configured at multiple levels (per API product, per developer app, or per user) and monitored in real time. In addition, Apigee

supports **traffic shaping, analytics, and automated quota resets**, giving administrators granular control over usage while ensuring reliable performance for all participants. Beyond access control, Apigee enables seamless **integrations** with internal and external systems, including support for versioning, mediation, and lifecycle management of APIs.

- **Connectivity:** Access for users (Developers, Academic institutions, Government departments, etc.) is secured via **Cloud Load Balancing** and **Cloud CDN**. Secure access to on-prem resources is managed via **Secure Hybrid Connectivity (VPN/VPC Connector)** and potentially **Cloud Interconnect** for high bandwidth.
- **API Usage Policies (initial defaults, tunable):**

  • **Burst control:** short-window throttles to protect shared services during spikes.
  • **Sustained rate-limits:** per API product and per developer app (e.g., per-minute ceilings).
  • **Daily quotas with automated reset** and **exception workflows** for approved research bursts.
  • **Tenant-aware keys/tokens** to ensure fair allocation across universities, startups, agencies.
  • **Real-time monitoring & alerts** (traffic, errors, latency) with audit trails of policy changes.

2. Control Plane & Governance

This is the non-negotiable security and audit layer that ensures compliance with the strict governance requirements (e.g., RBAC, NIST alignment) previously discussed.

- **Identity Management: Identity & Access Management (IAM)** enforces **SSO & Multi-Factor Authentication** (MFA). **Cloud IAM** implements **Role-Based Access Control (RBAC)** to manage access across the required data tiers: **Public, Restricted, and Internal**.
- **Security Perimeter:** The platform is secured using **VPC Controls** (VPC Service Controls perimeter) and the **Secret Manager** to protect sensitive credentials.
- **Auditing and Metrics: Cloud Logging** provides the mandatory audit functionality, ensuring **Consistent Identity, RBAC, Logging & Audit** across the system. A unique feature is the **Commons Credits Ledger**, which provides an **Auditable Transaction Log** for tracking **GPU/CPU Usage** and is architected to be **Blockchain-Ready** to support smart contract functionality.

3. Data Core & Curation Engine

This foundational layer is responsible for ingestion, metadata management, and structuring the data into a usable ecosystem. This addresses the RFP requirement for strong metadata management and moving beyond simple catalogs.

- **Knowledge Graph Core:** The data organization is centered on a **Knowledge Graph Core** that utilizes **Schema.org**, distinguishing it from older, dataset-centric catalogs (like CKAN/Dataverse) [172, 8.10, 55]. This enables inherent interoperability (FAIR principles).
- **Metadata Management: Data Catalog** is used for managing Metadata, Catalog functions, Alerts, and Governance.
- **Ingestion and Lineage: Data Ingestion Pipelines** utilize **Dataflow** for Batch/Scheduled Processing. The architecture supports the required **Dataset Lineage Tracking** and **Version Control** capabilities. The storage architecture employs a **Medallion Architecture** (layers of data quality).
- **Storage and Security:** Core storage uses **BigQuery** (Enterprise Data Warehouse) and **Cloud Storage** (Data Lake). **DLP (Data Loss Prevention)** is integrated to scan and protect sensitive data during ingestion and use.
- **Interoperability Standards:** The ingestion and modeling layer supports **healthcare interoperability** (e.g., **HL7/FHIR** schemas and vocabularies) and aligns with **W3C/Linked-Data** practices (e.g., **DCAT**, **Schema.org**) to advance FAIR principles across domains.

4. Central Storage & MLOps Platform (Zone A)

This layer provides the primary compute and storage resources for analytical workloads and serves as the command center for the hybrid architecture.

- **AI/ML Orchestration: Vertex** serves as the **Unified AI Platform**, responsible for the orchestration and control of resources in Zone B. It integrates with **BigQuery** and **Cloud SQL/Firestore (Commons Credits)**.
- **Responsible AI:** The platform integrates **Responsible AI** components, including **Vertex Model Eval**, which directly supports the RFP requirements for model fairness checks and continuous monitoring (as discussed in the history).

5. Zone B: Hosted/On-Prem - High-Performance Compute & AI

Zone B is architected as a highly secure, segregated environment focused on demanding workloads, aligning with the RFP's need for advanced, privacy-sensitive AI capabilities.

- **High-Performance Compute (HPC):** This zone hosts the **High-Performance Compute** resources, including **GPU/CPU Cluster Resources** in **Multi-tenant Sandbox Environments**. This is designed specifically to support external partners (Universities & Startups).
- **Synthetic Data Generation:** This critical RFP requirement is housed here. It supports **Multimodal Support** (Tabular/Text/Image/Video) and allows for fine-tuning **Privacy-Utility Tradeoffs**. It specifically employs leading open-source tools cited in the RFP, such as **SDV, Synthea, and CARLA**. Crucially, it must provide an **Audit Trail to Original Data** for compliance.
- **AI Fairness Tools:** Zone B hosts the specialized tools required for ethical AI development, including **Dataset & Model Fairness Checks** and **Bias Mitigation Tools**. These capabilities ensure **Continuous Fairness Monitoring** integrated into the ML Workflow, fulfilling the RFP requirements for fairness tools (like AIF360/Fairlearn).
- **Hardware & Hosting Profile (Zone B):** The DCC's high-performance workloads run at **MGHPCC/AICR** in a **secure, multi-tenant cluster** designed for elastic growth. The environment provides:
    - **GPU/CPU compute pools** for training/inference and synthetic data generation, provisioned via cluster schedulers with **per-tenant quotas** and isolation.
    - **High-throughput parallel storage** for short-lived training data and artifacts, plus policy-driven replication to long-term stores.
    - **Secure hybrid connectivity** (private links) to Zone A services for identity, catalog, logging, and credits metering.
    - **Initial capacity posture:** Sized for **hundreds of concurrent users** at launch with a roadmap to scale to **1,000+ concurrent users** through additional node pools and storage tiers; capacity reviews are part of quarterly governance.
    - **Operations note:** Hardware SKUs and final counts will be confirmed with **MGHPCC/AICR** and MassTech; the design supports gradual expansion without architectural change.

TOOL/SERVICE	LAYER	FUNCTION IN ARCHITECTURE	REASONING/RFP ALIGNMENT
BigQuery / Cloud Storage	Data Core & Central Storage	Provides the scalable **Enterprise Data Warehouse** and **Data Lake** infrastructure.	Essential for storing vast amounts of structured and unstructured data, supporting the required Medallion Architecture and data analytics workloads.
Data Catalog	Data Core	Centralized system for managing **Metadata, Catalog, Alerts, and Governance**.	Crucial for implementing metadata standards and achieving the **Findable** aspect of FAIR data principles. It forms the structure for the **Custom Data Commons**.
Dataflow	Data Core	Manages **Batch/Scheduled Processing** pipelines and ensures scalability to **Streaming** data.	Provides the scalable infrastructure required for the **Data Ingestion Pipelines**, supporting automated processing and ensuring data quality.
DLP (Data Loss Prevention)	Data Core	Used for integrating security and data scanning into the core pipeline.	Required to identify and protect sensitive data to comply with security standards (e.g., HIPAA, MA state privacy laws, as noted in the conversation history).
Cloud IAM	Control Plane & Governance	Implements **SSO, MFA**, and **Role-Based Access Control (RBAC)**.	Fulfills the foundational RFP requirement for robust **Security/Governance Frameworks**, ensuring data access is managed based on the user's role and data tier (Public/Restricted/Internal).

VPC Controls	Control Plane & Governance	Establishes a security perimeter around cloud services.	Crucial for preventing data exfiltration and maintaining the required level of security and boundary controls mandated by governance standards.
Apigee API Management	Presentation & Access	Manages APIs, **API Key/Token management**, and enforces **Rate limits**.	Provides secure external access, controlling and metering usage, which is necessary for managing resource consumption within the Commons Credits Ledger system.
Vertex (AI Platform) / Model Eval	MLOps & Central Storage	Acts as the **Unified AI Platform** (MLOps hub) and provides tools for model evaluation.	Provides the orchestration capability required to manage model training (controlling Zone B) and integrates essential **Responsible AI** tools, supporting **AI fairness** checks.
SDV, Synthea, CARLA	Zone B HPC	Specific tools for **Synthetic Data Generation**.	These tools are explicitly required (as illustrative examples) by the RFP to provide a robust, multimodal synthetic data capability, managing **Privacy-Utility Tradeoffs** in a dedicated secure zone.
TFDV (TensorFlow Data Validation)	Data Analytics & AI Orchestration	Likely used for proactive data analysis and validation within the pipeline.	Essential for pre-processing checks and detecting bias or anomalies early in the data preparation phase, supporting the required AI Fairness measures (similar to AIF360/Fairlearn referenced in history).

V. Enabling Financial Governance: The Commons Credits System

The DCC requires a central ledger to track "Commons Credits" to incentivize contributions, promote efficient usage, and sustain ecosystem engagement. This is implemented as a closed-loop incentive system, not a real-currency mechanism at launch, but can be expanded in future phases to include actual currency. Please refer to the below figure for a high level architecture for the commons credits approach.

Commons Credits Implementation Approach

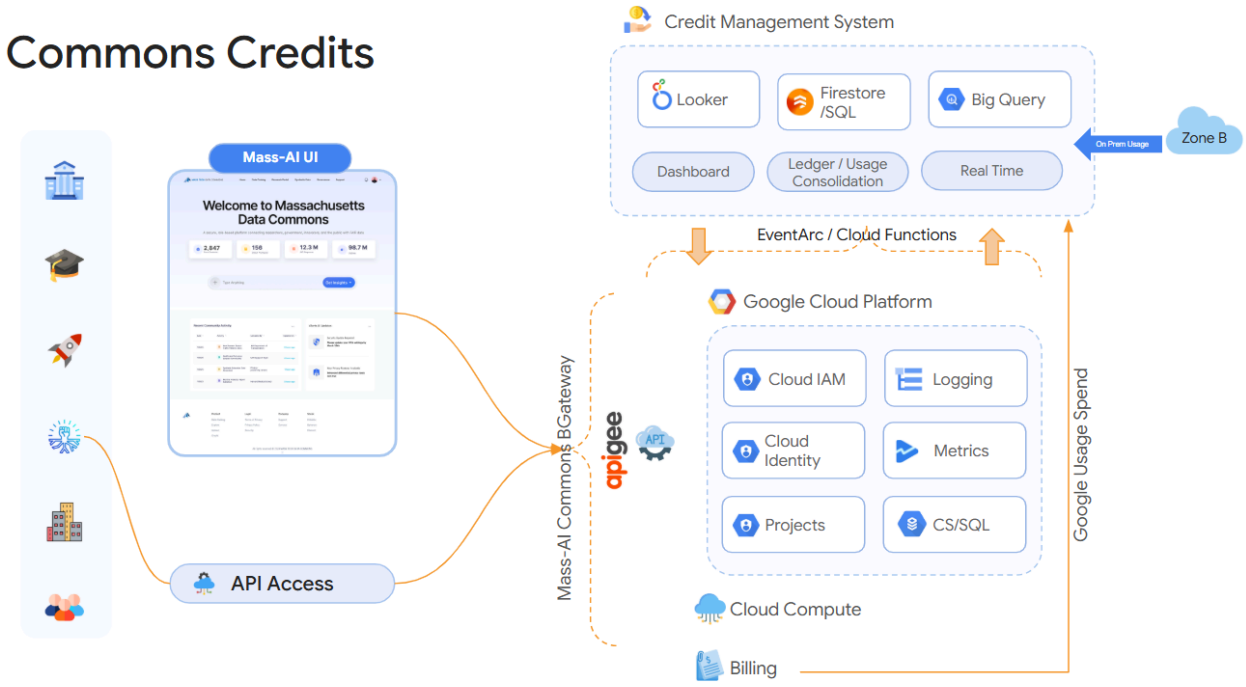
1. Unified API Access
    a. All usage flows through a single API layer managed by Apigee.
    b. The Mass-AI UI and programmatic clients consume the same APIs, ensuring consistent metering and policy enforcement.
2. Credit Services on Cloud Compute
    a. A dedicated Credit Management System runs as microservices on Cloud Run.
    b. Core endpoints include authorize, consume, refund, get-balance, and top-up.
3. Event Capture with Eventarc/Cloud Functions
    a. Every API call or backend job generates a usage event containing tenant, resource, and unit details.
    b. Eventarc or Cloud Functions routes these events to the credit processor for handling.
4. Ledger Implementation
    a. The **Cloud SQL database** serves as the **transactional system of record**, ensuring atomicity for balance changes and providing an auditable transaction history.
    b. All usage events, top-ups, refunds, and adjustments are also appended to **BigQuery** as an **immutable ledger**, consolidating internal Google Cloud usage (via Billing export) and external/on-prem usage into a single source of truth.
    c. The platform is designed to support **smart contracts** for automated enforcement of credit rules, and includes a **pluggable adapter layer** to support future blockchain integration without re-architecting. This balances audit integrity, usability, and flexibility.
5. Real-Time Balance in Firestore/SQL
    a. Current balances and entitlements are stored in Firestore or Cloud SQL for low-latency checks.

   b. The credit processor updates this store atomically, ensuring accurate real-time enforcement at request time.

6. Tracking & Reporting
   a. Credit usage logs (including API calls, data downloads, and compute resource consumption such as GPU/CPU usage) are streamed to **BigQuery**.
   b. **Looker dashboards** provide administrators and tenants with visibility into balances, burn-rates, usage patterns, and forecasts, supporting transparency and oversight needs.
   c. Batch reporting is sufficient initially, with options to evolve toward near real-time dashboards as adoption grows.
7. Billing and Reconciliation
   a. Google Cloud Billing data is exported to BigQuery and reconciled with the credits ledger.
   b. This enables cost recovery, grant allocation, anomaly detection, and margin analysis.
8. Payments and Top-Ups
   a. An **external payment processor** (e.g., Stripe, Adyen, or Kill Bill) handles monetary transactions.
   b. Successful payments trigger **TopUp events** that are recorded in BigQuery and reflected in Cloud SQL/Firestore balances.
   c. Peer-to-peer trading and real-currency purchases are **not in scope initially**.

9. Access and Integrity
   a. The system leverages **Cloud IAM** for secure authentication, with MFA required and role-based access control (RBAC) to protect sensitive operations.
   b. Cloud Audit Logs track all administrative changes, and immutable BigQuery facts plus WORM copies in Cloud Storage ensure tamper-evident integrity. The logs will act as a catch-all enabler for any custom events that need to be used for common credits computations.

10. Controls and Governance
    a. Quotas and rate limits are enforced at the API gateway.
    b. All financial events include idempotency keys to prevent duplication.
    c. Nightly recomputation of balances from the BigQuery ledger corrects drift and strengthens audit confidence.



Commons Credits High Level Architecture

VI. Differentiating AI Capabilities

Our integrated approach leverages Vertex AI to deliver superior capabilities for both synthetic data generation and ethical AI governance, which are critical components of the MA AI Hub vision.

VI.A. Synthetic Data Generation (RFP 2.4)

CLOUDSUFI moves beyond integrating standalone, domain-specific tools (SDV, Synthea, CARLA) by providing a **unified, multi-modal generative platform** centered on **Vertex AI**.

- **Technology: Gemini models** running on Vertex AI will generate high-quality, privacy-preserving synthetic data across diverse formats and modalities (including text, tabular, and multimodal data).
- **Workflow Integration:** The synthetic data pipeline is managed entirely within the **secure, governed Google Cloud environment** (BigQuery, Vertex AI Training). This eliminates risky data movement and simplifies MLOps.
- **Traceability and Validation:** The system supports schema-driven generation (e.g., using Snowfakery) and is required to maintain **audit trails linking synthetic datasets back to original real datasets** and validate configurable privacy-utility tradeoffs.
- **Target Domains:** Priority domains for synthetic data pilots include **lifesciences, robotics, education, and healthcare**.

VI.B. Responsible AI and Bias Framework (RFP 2.5)

We deploy **Google's comprehensive Responsible AI (RAI) toolkit**, embedding fairness, interpretability, and safety across the entire machine learning lifecycle.

- **Bias Detection in Data: TensorFlow Data Validation (TFDV)** will proactively profile training datasets to identify potential bias sources (e.g., data skew or under-representation) across protected attributes.
- **Model Evaluation and Debugging:** Post-training models will be evaluated using **Fairness Indicators** and the interactive **What-If Tool** to inspect performance metrics (e.g., accuracy, false positive rates) across sensitive features (e.g., race, gender, geography, socioeconomic factors).
- **Operationalized Monitoring: Vertex AI Model Evaluation** provides built-in data and model bias metrics, making fairness evaluation a first-class citizen in the MLOps pipeline. **Continuous monitoring** of fairness drift in deployed models is required.
- **Virtuous Cycle:** The integrated platform connects the synthetic data and fairness capabilities, allowing for the generation of synthetic data to augment under-represented minority groups and subsequent validation using TFDV and Fairness Indicators, all within a single, secure environment.
- **Reporting:** Fairness reporting is primarily for **internal auditing and compliance** in this phase, though it must enable public-facing transparency features in the future.

VI.C Deeply Embedded in Google Cloud's Innovation Efforts

CloudSufi is not an outside implementer but a trusted collaborator helping shape Google Cloud's native platforms. We are:

- Actively co-developing a governance UI layer on top of Google Cloud Dataplex, integrating metadata, lineage, and quality APIs for real-time visibility and policy enforcement.
- Engaged in the multi-year Data Commons program, having been awarded the three-year Data Commons Data Loading Statement of Work (SOW) — underscoring our role in expanding and operationalizing Google's Data Commons knowledge graph.
- Focused on extensibility and scale, ensuring that Massachusetts benefits not only from proven Google technologies (Data Commons, Dataplex, BigQuery, Apigee, Vertex AI, Looker) but also from CloudSufi's direct, ongoing contributions to these ecosystems.

This close alignment with Google Cloud means the Massachusetts AI Hub is not adopting experimental technology or unproven tools; it is leveraging the same operational infrastructure and expertise that powers Google's innovation roadmap.

Key Scope & Accomplishments

- Designing and developing a user-friendly governance UI layer on top of Google Cloud Dataplex.
- Integrating Dataplex metadata, lineage, and data quality APIs for real-time visibility into data assets.
- Enabling policy management, data cataloging, and compliance workflows directly within the UI.
- Automating metadata synchronization between Dataplex and the governance platform.
- Supporting governance at scale across multiple domains and distributed environments.
- Delivering a scalable, extensible solution adaptable to evolving enterprise needs.

Business Benefits

- Simplified Data Governance – intuitive interface for policy management and compliance.
- Improved Visibility – clear insights into data lineage, quality, and asset health.
- Enhanced Collaboration – enables stakeholders to govern and share data seamlessly.
- Faster Decision-Making – accelerates access to trusted, governed data.
- Scalability – designed for large, distributed data environments without complexity.

VII. Security, Governance, and Compliance Framework

Security is implemented through a multi-layered, defense-in-depth posture native to GCP, ensuring compliance with state requirements. **Audit Retention & Archival: ≥3 years** log retention at launch with an **archival strategy to be finalized** with MassTech. Logs are exportable to SIEM (e.g., Splunk/ELK) for security operations and compliance reporting

SECURITY & GOVERNANCE REQUIREMENT	CLOUDSUFI SOLUTION (GCP COMPONENT)	REQUIREMENT DETAIL/CITATION
Access Control & Identity	Cloud IAM / Identity Platform	Granular **Role-Based Access Control (RBAC)** and support for **Multi-Factor Authentication (MFA)** are mandatory. Supports multi-tiered access: Public, Restricted (requires review), and Internal.
Dataset Tiering & Security	Cloud Data Catalog Policy Tags	Enables fine-grained, **column-level security** within BigQuery to control access to sensitive information. Adheres to data classification security standards set by the MA CISO.
Data Protection	Default Encryption (CMEK Option)	All data is **encrypted by default** at rest (in GCS/BigQuery) and in transit. Supports the use of Customer-Managed Encryption Keys (CMEK) for Commonwealth key authority.
Network Perimeter	VPC Service Controls	Establishes a virtual security boundary, preventing data exfiltration even if credentials are compromised.
Auditing and Logging	Cloud Audit Logs / SIEM Integration	Provides a complete, immutable record of all administrative actions and data access events. Audit logs must be retained for a minimum of **3+ years**. The system is designed for integration with SIEM tools (Splunk, ELK).

Data Residency	US Regions (East-coast Preference)	All data (including logs and model artifacts) will be stored and processed in **U.S. regions** with an East-coast preference. **Remote access** to Commonwealth Data from outside the continental United States is **prohibited**.
Consent Management	Cloud SQL / Integrated Workflow	Tokenized, revocable, and time-bound consent management is **in scope for Phase 1** for restricted datasets.
Governance Alignment	NIST/CISA/Applicable Laws	Governance adheres to applicable state and federal laws. Vendors are encouraged to align with recognized standards like the **NIST AI Risk Management Framework**. Defining governance policies will be a **collaborative effort** with MassTech.

VIII. Project Management, Delivery, and Ecosystem Growth

CLOUDSUFI (via Google Cloud PSO) will lead the implementation using a proven methodology, focusing on transparent program management, stakeholder coordination, and phased delivery. The anticipated start date is **December 2, 2025**. A **phased release** approach is preferred to achieve early usability.

1. Project Delivery & Environments

   We set up **Dev/Test, QA, and Production environments**. The delivery plan includes rigorous QA testing, tailored training programs for end-users and administrators, and comprehensive documentation.

2. User Experience & Accessibility

   The platform must meet **WCAG 2.1 Level AA accessibility standards** at a minimum. The architecture will support **multilingual accessibility** (e.g., screen reader compatibility for non-English content) for future localization.

   Accessibility QA: We will validate **WCAG 2.1 AA** with **screen reader compatibility**, **keyboard-only navigation**, and **captioning/transcripts** for multimedia. Testing includes assistive-technology checks and UAT with accessibility criteria, with results documented for acceptance.

3. Data Curation and Integration

   We will collaborate with MassTech to refine the initial dataset strategy and propose the initial set of datasets, drawing from the illustrative list (Appendix A). Integration will use **proven, open-source Data Commons import tools** that map local data into the knowledge graph's schema. We will leverage APIs for integration with external state systems (e.g., MassGIS, MassDOT).

4. Ecosystem Engagement (RFP 5.3)

   We are committed to fostering the local ecosystem.

   - **Academia:** Provide **Google Cloud research credits** and deliver workshops/training modeled on existing Data Commons educational courseware.
   - **Startups:** Offer Cloud credits, mentorship, and business support through the **Google for Startups Cloud Program**.
   - **Government:** Collaborate with EOTSS to integrate the DCC's queryable knowledge graph capabilities with the existing Massachusetts Data Hub.

Total Estimated Duration (Kickoff to Public Launch): **40 Weeks** (Accelerated Phased Release)

CLOUDSUFI proposes an **accelerated 40-week phased release** schedule, leveraging the inherent efficiencies and pre-built components of the **Google Data Commons architecture on GCP**. This approach directly addresses the RFP's preference for a phased release and feasibility of timeline.

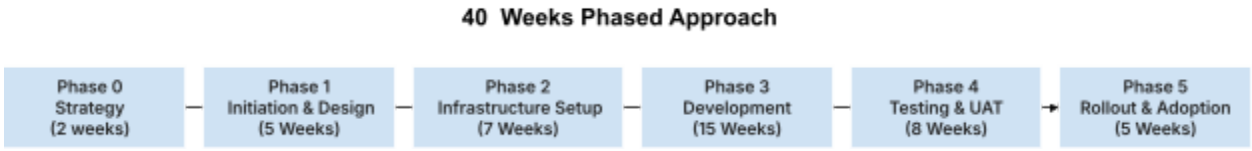
The plan is structured around the five mandated Program Phases outlined in DCC Requirements Section 3.0. We prioritize the implementation of the core foundation (Phase 2), including the integrated **Cloud Data Catalog** and the **Commons Credits ledger**, followed immediately by the deployment of differentiating AI capabilities (**Vertex AI for Synthetic Data** and the **Responsible AI (RAI) Toolkit**) in Phase 3. This ensures the platform is secure, functional, and equipped with high-value AI features for initial pilot users well ahead of the final public launch.

Key Acceleration Drivers

1. **Platform Instantiation:** We avoid custom-building the knowledge graph, instead instantiating a customized, secure version of the **Google Data Commons knowledge graph**, significantly accelerating the Data Integration phase.
2. **Native Integration:** Utilizing tightly integrated GCP services (BigQuery, Vertex AI, Cloud Data Catalog) eliminates time lost integrating fragmented open-source tools mentioned in the RFP.
3. **Concurrent Development:** We run website, metadata, and AI feature development concurrently within the mandated **Dev/Test and QA environments**.

Detailed Project Plan (40 Weeks)

The plan outlines phased delivery of MA AI Hub Data Commons Collaborative, through a phased and sprint-based delivery approach, ensuring alignment with business objectives, technical readiness, and governance requirements. The plan is structured into five phases, with activities running in parallel to optimize delivery speed and value realization:



Phase 0

Investment -Strategic Assessment (2 Weeks): CLOUDSUFI, supported by Google Cloud, will dedicate an expert Assessment POD to run a rapid 2-week discovery using our proven **Data Commons Readiness Framework**. This phase aligns key stakeholders, validates assumptions, and surfaces strategic opportunities before the build begins. Deliverables include a readiness baseline, risk map, and high-level blueprint that de-risks execution and accelerates decision-making. As an **investment** from our side, this ensures MassTech enters Phase 1 with clarity, consensus, and a focused roadmap

Phase 1

Initiation & Design (5 Weeks): Project kickoff with team alignment and charter approval. Finalize detailed architecture and governance strategy. Review and approve dataset curation criteria and confirm the initial dataset list.

Phase 2

Infrastructure & Setup (7 Weeks): Provision cloud infrastructure, identity management, and foundational services. Validate development and test environments to support synthetic tooling and backend deployment.

Phase 3

Development (15 Weeks): Complete initial data ingestion pipelines and metadata catalog. Operationalize the synthetic data engine to produce test outputs. Deliver a functional website with browsing, search, and dataset detail pages. Deliver an admin dashboard for dataset management.

Phase 4

Testing & Governance (8 Weeks): Conduct platform QA across functional, performance, and security dimensions. Implement the governance framework, including RBAC, data tiers, and usage agreements.

Phase 5

Rollout & Adoption (5 Weeks): Finalize user documentation and deliver tailored training. Onboard initial user groups. Gather pilot feedback and prepare the platform for public launch and ecosystem adoption.

This 2 weeks of sprint-driven approach ensures that platform infrastructure, AI capabilities, and governance frameworks evolve in tandem, supporting a robust, secure, and scalable delivery

SPRINT	WEEKS	KEY ACTIVITIES (PARALLEL WORKSTREAMS)	DELIVERABLES / OUTCOME	DCC REQUIREMENT ALIGNMENT
Sprint 1	W1–W2	Initiation: Project Kickoff, core team resourced and aligned.	Project Kickoff, Architecture & Governance Defined	Project Initiation & Governance Policies
Sprint 2	W3–W4	Design: Finalize high-level Architecture & Governance strategy, define initial governance policies (ethical use, data tiering).	Architecture & Governance Defined	Project Initiation & Governance Policies
Sprint 3	W5–W6	Design: Refine and gain approval for Dataset Curation Criteria, finalize proposed data sourcing strategy and data types (tabular, text, image, etc.).	Dataset Curation Criteria Approved	Dataset Curation & Sourcing
Sprint 4	W7–W8	Design: Validate and confirm Initial Dataset List for hosting and external linking.	Initial Dataset List Confirmed	Dataset Curation & Sourcing
Sprint 5	W9–W10	Infrastructure: Platform Infrastructure Provisioned (Production environment setup: GCS, BigQuery, VPC Service Controls perimeter).	Platform Infrastructure Provisioned	Secure GCP Foundation & Access Management
Sprint 6	W11–W12	Infrastructure: Configure Cloud Identity & Access Management (IAM), define RBAC roles, implement Cloud Audit Logs.	IAM Configured	Secure GCP Foundation & Access Management

Sprint 7	W13–W14	Infrastructure: Dev/Test Environments Ready, deploy Data Commons backend (Cloud Run, Cloud SQL), configure AI Compute Resources (AICR).	Dev/Test Environments Ready	Dev/Test Environment Setup & Data Ingestion
Sprint 8	W15–W16	Development: Build initial Data Ingestion Pipelines using Data Commons import tools (CSV/MCF mapping), support non-standard formats.	Data Integration Started	Dev/Test Environment Setup & Data Ingestion
Sprint 9	W17–W18	Infrastructure: Deploy Google Cloud Data Catalog, implement automated metadata ingestion from BigQuery and GCS. Governance: Commons Credits Ledger Setup (Phase 1), implement Column-Level Security using Data Catalog Policy Tags.	Metadata Catalog Setup, Commons Credits Phase 1	Metadata Management & Governance Tools
Sprint 10	W19–W20	Development: Finalize Initial Data Catalog and Metadata System, integrate Cloud DLP for automated PII tagging, begin semantic search functionality.	Metadata Catalog Functional	Data Access, Privacy, and Website Functionality
Sprint 11	W21–W22	Development: Synthetic Data Engine Operational – provision Vertex AI environment, configure Gemini models. Website: Core functional browsing, basic search, filtering (WCAG 2.1 AA compliant).	Synthetic Data Engine Setup, Website Core Functional	Data Access, Privacy, and Website Functionality

Sprint 12	W23–W24	Development: Generate synthetic data test outputs for priority domains, validate configurable privacy-utility tradeoffs. Website: Implement Dataset Detail Pages, integrate Gemini in Looker for natural language querying.	Synthetic Data Engine Operational, Website Enhancements	Synthetic Data & Website Enhancements
Sprint 13	W25–W26	AI Capabilities: AI Fairness Framework Integration – integrate TFDV for data bias profiling, configure model evaluation hooks. Admin: Develop Admin Dashboard UI for managing submissions and user suggestions, implement multi-tenant administration capabilities. Data Integration: Complete pipelines for initial hosted datasets, configure version control and provenance tagging.	AI Fairness Setup, Admin Dashboard Live, Data Integration Complete	AI Fairness, Admin Controls & Provenance
Sprint 14	W27–W28	API Development: Finalize API-based data access (REST/Python APIs), implement API Gateway (rate limiting, API keys). Commons Credits: Finalize Sample Commons Credits ledger output, dashboard integration.	APIs Functional, Credits System Finalized	API Access & Commons Credits Integration
Sprint 15	W29–W30	Development: Finalize all core Website UI/API features, data integration, AI fairness, and credits system. Testing: Prepare for Internal QA of the end-to-end system.	Dev Deliverables Complete, QA Ready	QA Preparation & System Readiness

Sprint 16	W31–W32	Testing: Internal QA Testing – end-to-end testing of core functionality, synthetic data pipeline, and API access. Infrastructure: Performance tuning and load testing.	Platform QA, Performance Tested	Internal QA & Performance Validation
Sprint 17	W33–W34	Security: Independent security review against MA CISO standards, address critical findings. Governance: Fairness & Bias Reporting Dashboard operational, Governance Framework operational (tokenized, revocable, time-bound consent for restricted datasets).	Security Review Complete, Governance Framework Operational	Security & Governance Enforcement
Sprint 18	W35–W36	Training & Rollout: Finalize training and launch materials, tailored training for admins, researchers, and non-technical users. User Onboarding: Deliver training to initial administrators and pilot user groups, initiate pilot feedback collection. Ethical AI Training & Certification: Establish role-based training (researchers, admins, developers) covering responsible data use, bias detection, and RAI practices, with completion tracking and attestation prior to elevated access.	Training Delivered, User Onboarding & Pilot Feedback	User Training & Pilot Engagement
Sprint 19	W37–W38	Documentation & Engagement: Finalize Detailed Platform Architecture Documentation, formalize planned industry engagement	Documentation & Engagement Complete, Platform Stabilized	Documentation & Industry Engagement

		(researchers/credits, startups, EOTSS/Mass Data Hub integration). Public Launch Prep: Prepare for public launch, system stabilization.		
Sprint 20	W39–W40	Rollout & Post-Launch Support: Public Launch of MA AI Hub Data Commons Collaborative, monitor adoption, address critical post-launch issues, finalize knowledge transfer to operations.	Platform Stabilized, Knowledge Transfer Complete	Public Launch & Post-Launch Support

Project Governance & Change Management

Effective governance and structured change management are crucial for a project's success. This approach ensures the project delivers its intended outcomes while remaining adaptable to new requirements and unforeseen challenges.

Governance Approach

A clear governance framework will guide strategic decisions and maintain accountability -

- **Project Steering Committee:** This committee, comprising client and delivery leadership, will provide high-level oversight, resolve major issues, and make strategic decisions to keep the project on track.
- **Project Management Office (PMO):** The PMO will handle the day-to-day management of the project, including monitoring risks, tracking progress against milestones, and ensuring all activities comply with governance policies.
- **Defined Roles and Responsibilities:** Roles and responsibilities will be clearly defined and documented for all project stakeholders to ensure transparent decision-making and clear accountability.

Change Management Process

A disciplined change management process will handle changes efficiently and with full transparency -

1. **Change Identification:** Any proposed changes to the project's scope, new requirements, or unexpected issues will be formally submitted through a **change request process**.
2. **Impact Assessment:** Each request will undergo a thorough assessment to evaluate its impact on the **project's scope, schedule, cost, resources, and risks**.
3. **Approval Workflow:** The Steering Committee will review each request and either approve or reject it. The client will be directly involved in the approval of all material changes.
4. **Controlled Implementation:** Once approved, the change will be integrated into the project's baseline plan. This includes updating deliverables, timelines, and communication plans to all relevant stakeholders.

The structured approach ensures the project stays aligned with the priorities, while providing the discipline and transparency needed to manage changes without disrupting overall delivery.

RFP Requirements Compliance Checklist (Phase Integration)

The following checklist confirms that all required services and deliverables (RFP Section 2) are addressed within the proposed 40-week phased plan.

RFP SERVICE AREA (RFP SEC. 2)	RFP DELIVERABLE (RFP SEC. 2.8)	PHASE WHERE CORE IMPLEMENTATION OCCURS	CLOUDSUFI SOLUTION COMPONENT
1. Program Management & Coordination	Detailed Project Plan and Timeline	Phase 1 (W1) & Phase 5 (W36)	PSO Methodology & Integrated Project Plan
2. Technical Architecture & Platform Development	Detailed Platform Architecture Documentation	Phase 1 (W5) & Phase 5 (W36)	GCP Serverless Architecture (BigQuery, Cloud Run, Cloud SQL)
	Initial Data Catalog and Metadata System	Phase 2 (W9) & Phase 3 (W16)	**Google Cloud Data Catalog** (Automated sync, PII detection)
	Sample Commons Credits ledger output	Phase 2 (W10) & Phase 3 (W28)	Cloud SQL Ledger, BigQuery/Looker Reporting
3. Dataset Curation and Integration	(Integrated into plan)	Phase 1 (W4) & Phase 3 (W26)	Data Commons Knowledge Graph & Import Tools
4. Synthetic Data Capability	Synthetic Data Tooling and Sample Outputs	Phase 3 (W18 - W19) & Phase 4 (W32)	**Vertex AI with Gemini Models**

5. AI Fairness and Bias Framework	Fairness and Bias Reporting Dashboard	Phase 3 (W25) & Phase 4 (W31)	Google **Responsible AI Toolkit** (TFDV, Fairness Indicators)
6. Security, Governance & Compliance	Security and Governance Implementation	Phase 2 (W6 - W12) & Phase 4 (W33)	Cloud IAM, **VPC Service Controls**, Cloud Audit Logs
7. Testing, Rollout, and Adoption	Training and Launch Materials	Phase 4 (W30) & Phase 5 (W34)	Rigorous QA, Tailored Training Programs
8. Deliverables	Pilot user feedback	Phase 5 (W38)	Structured feedback collection from MassTech-recruited users
	Explanation of planned industry engagement	Phase 5 (W40)	Plan for Researchers (credits), Startups (programs), EOTSS/Mass Data Hub

IX. RFP Requirements Compliance Checklist

This table maps the RFP requirements and deliverables (Section 2) to the proposed CLOUDSUFI (Google Cloud) solution components, confirming compliance and indicating areas where the proposed solution exceeds requirements.

RFP SECTION	REQUIREMENT/DELIVERABLE	PROPOSED CLOUDSUFI SOLUTION & APPROACH	MEETS/ EXCEEDS
2.1 Program Management & Coordination	Integrated project plan, regular meetings, milestone tracking	Google Cloud PSO methodology leveraging deep collaboration and transparent communication.	Exceeds
2.2 Technical Architecture & Platform Development	Design/build website and backend infrastructure	Serverless architecture on GCP using BigQuery, Cloud Run, Cloud SQL, and Cloud Storage, based on proven Data Commons reference architecture.	Exceeds
	Implement metadata management, cataloging, and discovery tools (e.g., CKAN, Dataverse)	**Google Cloud Data Catalog** with automated ingestion, rich structured tags, and Google-powered search.	Exceeds
	Develop role-based access and identity management system	**Google Cloud IAM** and Identity Platform providing robust RBAC, MFA, and delegated administration.	Meets
	Develop a central ledger for tracking "Commons Credits"	Complete system using **Cloud SQL** (transactional ledger), **Cloud IAM** (identity), and **BigQuery/Looker** (reporting), designed for smart contracts and future pluggable backends.	Exceeds
	Enable API-based data access and usage analytics	Core Data Commons REST and Python APIs enabled. Usage analytics captured in BigQuery and reported via Looker	Meets

		dashboards. Rate limiting and API key management implemented.	
2.3 Dataset Curation and Integration	Assist in curating initial datasets	Collaborative curation using proven, open-source Data Commons import tools for harmonization into the knowledge graph.	Meets
	Ensure metadata completeness and FAIR data principles	Knowledge graph architecture is a native, mature implementation of FAIR principles.	Exceeds
	Support ingestion pipelines, version control, and provenance tagging	Data Commons import process includes provenance tagging and dataset-level version control.	Meets
2.4 Synthetic Data Capability	Integrate synthetic data tools (e.g., SDV, Synthea, CARLA)	Integrated synthetic data generation using **Gemini models on the Vertex AI platform**, supporting multi-modal data types within a governed MLOps environment.	Exceeds
2.5 AI Fairness and Bias Framework	Implement bias detection, auditing, and mitigation tools (e.g., AIF360, Fairlearn)	**Google's comprehensive Responsible AI (RAI) toolkit**, including TFDV, Fairness Indicators, What-If Tool, and native Vertex AI Evaluation metrics for continuous monitoring.	Exceeds
2.6 Security, Governance & Compliance	Define and implement RBAC, MFA, and dataset tiering; ensure logging, encryption, and audit trails	Multi-layered security including Cloud IAM, VPC Service Controls, default encryption (CMEK option), Cloud Audit Logs (3+ year retention), and Phase 1 tokenized consent.	Exceeds
2.7 Testing, Rollout, and Adoption	Conduct QA, train end users/administrators, prepare	PSO-led delivery including QA, tailored training, and comprehensive documentation for technical and	Meets

	documentation, support public launch	non-technical users. Compliance with WCAG 2.1 AA.	
2.8 Deliverables	Detailed Platform Architecture Documentation	Full GCP architecture documentation provided (Deliverable).	Meets
	Fairness and Bias Reporting Dashboard	Dashboards built in Looker and Vertex AI, leveraging Fairness Indicators and model evaluation outputs for internal auditing.	Meets
	Explanation of planned industry engagement	Detailed plan for engaging researchers (Cloud Credits), startups (Google for Startups), and government agencies (EOTSS integration).	Meets

X. Assumptions

The CLOUDSUFI project plan and technical architecture are based on the following key assumptions regarding the execution environment, data availability, and governance structure:

1. Architectural and Technical Assumptions

- **Hybrid Deployment Model Confirmed:** The solution will utilize a **hybrid model**, leveraging **MGHPCC/AICR** for hosted data and compute, paired with **cloud-based services** (e.g., Google Cloud Platform, as proposed by CLOUDSUFI) for the catalog, APIs, and web access.
- **API Availability for Integration:** Vendors can assume **APIs will be made available** for integration with external Massachusetts systems, such as MassGIS, MassDOT, and Public Health data sources.
- **Infrastructure Procurement:** MassTech will be responsible for **procuring the selected infrastructure independently**, including necessary licenses and hosting platforms; these are **not pass-through costs** for the vendor.
- **Environment Requirements:** Three distinct environments are required for delivery: **Dev/Test, QA, and Production**.
- **Data Ingestion Approach: Batch processing will be sufficient** for data ingestion in Phase 1, with real-time or streaming capabilities deferred to future phases.
- **Data Volume Estimates:** Initial data volumes are anticipated to be in the **tens of TBs with approximately 30% annual growth**.

2. Data and Scope Assumptions

- **Initial Dataset Scope:** The datasets referenced in DCC Requirements Appendix A are **illustrative examples and are not mandatory inclusions**; scope includes finalizing a recommendation for # datasets should be sourced & integrated during the initial launch.
- **Curation Responsibility:** The respondent is expected to **identify and propose high-value datasets** for inclusion, working in partnership with MassTech.
- **Pilot User Recruitment:** MassTech will **identify and recruit pilot users** from the stakeholder community, and the vendor will collaborate on the outreach.
- **Security Standard Alignment:** The DCC will adhere to the **data classification security standards** set by the MA CISO. Governance must also comply with applicable state and U.S. federal laws, and vendors are encouraged to align with standards such as the **NIST AI Risk Management Framework**.

3. Governance and Personnel Assumptions

- **Onshore Requirements:** Key project functions involving **governance, security, compliance, and stakeholder engagement** must be performed by **U.S.-based personnel**.
- **Policy Definition:** Defining governance policies and standards (e.g., ethical usage, bias detection, data tier delineation) will be a **collaborative effort** with MassTech and its established governance committee.
- **Start Date:** The anticipated project start date is **December 2, 2025**.

4. Out-of-Scope Items (Exclusions and Future Phase Deferrals)

The following activities, functionalities, and requirements are explicitly considered out of scope for the current phase of the project:

4.1 Commons Credits and Financial Systems

- **Real-Currency Transactions:** The Commons Credits system will operate as a **closed-loop incentive system** within the DCC, **not a real-currency mechanism**. Real-currency purchases are **out of scope initially**.
- **Credit Trading and Currency: Peer-to-peer credit trading** between users and **multi-currency support** (e.g., credits to USD exchange) are **not in scope**.
- **External Financial Integration:** Integration with **COMMBUYS** or performing **accounting and invoicing for cross-agency cost allocation** is **out of scope**.
- **Federal Grant Integration:** Integration with federal funding sources (e.g., NIH, NSF grants) for research usage tracking is **out of scope** for this phase.

4.2 Data Access, Storage, and Operations

- **Offshore Data Access: Remote access to Commonwealth Data from outside the continental United States** is strictly **prohibited** unless approved in advance in writing by the Commonwealth.
- **Vendor Data Storage:** The contractor is **not responsible for storing any data** in the contractor's computing environment.
- **24/7 Operational Support:** Ongoing **24/7 system support** is **not in scope** for this RFP; the vendor will provide support through the initial rollout phase.
- **Streaming Data: Real-time feeds and streaming datasets** (e.g., Kafka/Event Hub integration) are **out of scope** for the initial phase of work, planned for future releases (Year 2 or 3).

- **Offline/Low-Bandwidth Modes:** Support for **offline or low-bandwidth modes** for mobile users is **not in scope** for this phase.
- **Archival Strategy:** An archival strategy for logs **beyond three years** has not been defined.

4.3 AI and Fairness Capabilities

- **Real-Time Bias Monitoring: Real-time monitoring** of deployed models for fairness drift is **not required** for this phase; the architecture should include hooks for future capability.
- **Fairness Reporting for Public: Public-facing fairness dashboards** are **not required** for the initial phase; reporting is primarily for internal auditing and compliance.
- **Interactive Scenario Testing: Interactive scenario testing** (e.g., What-If Tool functionality) is **not in scope** for this phase.
- **Synthetic Data Production Grade:** Synthetic data pipelines are intended primarily for **R&D and sandbox environments**, not full production-grade usage.

4.4 Governance and Project Scope

- **External Site Access:** Access to **other MassTech sites and platforms** (e.g., Single Sign-On for sites outside the DCC) is **out of scope**.
- **Third-Party Certification:** A formal **third-party accessibility audit or certification** is **not mandated** as a baseline requirement in the RFP.
- **Multilingual UI at Launch:** While the platform must meet WCAG 2.1 AA accessibility standards and the architecture must enable future localization, **multi-language support** (English + others) is **not strictly required in Phase 1**.
- **Mandated Technologies:** The tools listed in the RFP (e.g., CKAN, Dataverse, SDV, Synthea, AIF360, Fairlearn) are **illustrative examples**, and **not mandatory components** of the technology stack. Vendors are free to propose alternatives.

XI. Past Performance & References

Case Study 1
Knowledge Fabric for Technology Disruption Framework

CONTACT PERSON	Chief Technology officer (CTO)
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

CLOUDSUFI successfully provided Data Engineering Services for Knowledge Fabric, supporting hedge fund's AI's product data consumption and ongoing data updates. Services, primarily in the US and London with remote support, established a robust foundation for research data, focusing on the Enterprise Knowledge Graph with Validation Agent.

Scope & Accomplishments

Data Corpus enrichment: Identified and validated data sources (BNEF, World Economic data, IRENA). Developed automated 1000+ pipelines and ETLs. Built the Stellar Knowledge Fabric using a Graph Database for an Enterprise Knowledge Graph.

Data Validation agent

Created, trained, and deployed a Validation Agent model to ensure data accuracy. Digitized all interconnected data assets within the Knowledge Fabric.

Business Benefits

- Enhanced Data Accuracy and Reliability: Validation Agent ensures high data quality.
- Accelerated Investment Decisions: Enterprise Knowledge Graph provides comprehensive market data for quicker, informed choices.
- Operational Efficiency: Automated processes streamline data handling, reducing research timelines by 90%.

Case Study 2
Patient 360 Platform

CONTACT PERSON	Chief Information officer (CIO)
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

Overview

CloudSufi, with GCP and DataCommons.org, built a **FAIR-compliant Patient 360 Platform** for a U.S. healthcare provider. The platform unified fragmented patient data (EMR, claims, labs, pharmacy, wearables, and social determinants) into a secure, AI-ready environment.

Challenge

Data silos, poor metadata, compliance risks (HIPAA), and limited AI capabilities prevented holistic patient insights and slowed research.

Solution

- **Platform:** Multi-tenant data lake and metadata catalog on GCP with DataCommons.org schema.
- **AI/ML:** Predictive risk scoring, outcomes forecasting, and synthetic data generation with Vertex AI.
- **Governance:** HIPAA-compliant security, audit trails, and fairness/explainability frameworks.

Impact

- 15% reduction in readmissions.
- 20% cut in admin overhead via automation.
- 90% decrease in reporting errors.
- FAIR-compliant catalog enabling faster research (minutes vs. weeks).

Case Study 3
Industry : Fashion and Jewelry

CONTACT PERSON DESIGNATION	Director of Data Engineering & AI
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

Challenge

- **Challenge 1:** Fragmented Data Landscape, data trapped across multiple, disconnected legacy ERP systems
- **Challenge 2:** Rigid, monolithic legacy ERP architecture was unable to support modern needs, limiting capabilities for real-time analytics and AI applications

Impact

- **Established Single Source of Truth:** Successfully implemented 1000+ data objects on BigQuery, creating a trusted, consistent foundation for all forward-looking business decisions and enterprise reporting
- **Accelerated Migration with AI:** Leveraging our in-house AI agents to automate complex transformation logic and mapping, reducing the time and manual effort traditionally required for large-scale data migration

Business Benefits

- **Empowered Strategic Decision-Making:** Shifted business reporting from retrospective analysis to forward-looking, scenario-based insights, ensuring leaders can make proactive decisions that directly impact revenue and operational efficiency

Case Study 4
Data Lake Modernisation

CONTACT PERSON	Senior Product Manager (Google Cloud)
EMAIL	To be provided upon request
PHONE NUMBER	To be provided upon request

Challenge

- **Challenge 1:** Customers had to develop their own custom solution to ingest data into BigQuery and also to publish the outbound data to external target systems. This challenge was addressed by standard managed connectors developed to cater to all the critical Target systems.
- **Challenge 2:** Templatized, intuitive, simplified UI based configuration driven connectors to connect to any source/sink to/from Google BigQuery

Impact

- **High-Performance Data Integration Connectors:** Developed SaaS connectors with 100% object coverage, industry-leading throughput, continuous enhancements based on enterprise feedback, and comprehensive maintenance/implementation guides.
- **Robust Data Validation Framework:** Created a specialized Test Framework for complex ETL data integrity, handling intricate data type mappings, and implementing CI/CD for automated validation.

Business Benefits

- **Seamless Third-Party Integrations:** Ensures high data quality and throughput for complex SaaS ETL (batch and real-time).
- **Guaranteed Data Integrity:** Our Validation Framework simplifies ETL maintenance and ensures data accuracy via periodic upgrades.
- **Enhanced Operational Efficiency:** Implementation guides significantly reduce customer overhead during large deployments.

XII. Bios of Key Personnel

#	NAME	ROLE TITLE	KEY RESPONSIBILITIES	REQUIRED SKILLS & TECHNOLOGIES
CLOUDSUFI				
1	Irfan Khan	Executive Sponsor	Provide executive oversight, ensure alignment with MassTech leadership, facilitate escalation and partnership with Google Cloud	Executive leadership, Strategic Partnerships, Stakeholder Engagement
2	Walid Hassan	Technical Program Manager	Lead program planning and execution, manage timelines and deliverables, run Agile ceremonies, provide risk and change management, ensure stakeholder communication.	Agile, Scrum/SAFe, Jira, Confluence, Risk Management, Cloud Program Delivery
3	Chiranjeev Dahiya	Business Analyst	Engage stakeholders to capture business and functional requirements, translate needs into user stories, validate against FAIR data principles, support testing documentation	Requirements Gathering, Agile, JIRA, Metadata Standards, FAIR Principles, Documentation
4	Vasudev Nayak	Cloud Solution Architect	Design and oversee end-to-end DCC architecture on Google Cloud; ensure scalability, security, and compliance; guide technical teams on integration patterns.	Google Cloud Architecture (BigQuery, Dataflow, IAM, Vertex AI), Security & Compliance, API Management
5	Amit Chakraborty	Integration Expert	Build and maintain secure data ingestion pipelines, integrate metadata catalog and synthetic data tools, ensure seamless interoperability with APIs and external datasets.	Workflow Orchestration (Dataflow, Airflow), Application Integration, CKAN/Dataverse, Synthetic Data Tools (SDV, Synthea, CARLA), API Management

Google Cloud				
7	Annette Ball	Account Executive - Google Public Sector	Serve as Google's executive liaison to MassTech; ensure access to the right Google resources; facilitate contract alignment, licensing, and strategic initiatives for public sector engagements.	Public Sector Account Management, Strategic Partnerships, Google Cloud Contracts & Licensing, Stakeholder Engagement
8	Mina Behnam	Senior Customer Engineer Google Cloud	Provide deep technical expertise on Google Cloud services; support solution architecture, security/compliance mapping, and performance optimization; ensure MassTech benefits from best practices and Google's innovation roadmap.	Google Cloud Architecture (BigQuery, Vertex AI, IAM, Dataflow), Security & Compliance (FedRAMP, HIPAA), Data & AI Best Practices, Cloud Cost Optimization

Line Number	Budgeting																	
2	Applicant Information																	
3	Applicant: CLOUDSUFI Inc			Procurement No.:														
4	Budget Period:			Title of Proposed Project:														
5	Address: 333 W San Carlos St, San Jose, CA 95110			AI Hub Data Commons Collaborative														
6																		
7				Consulting Services														
8				MTC Funding Requested:														
9				Total Project Cost:														
10	Cost Elements				Total Project	Invoice 1		Invoice 2		Invoice 3		Invoice 4		Cumulative				
11						Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY				
12							Amount		Amount		Amount		Amount	Expended		Remaining		
13	I. Direct Labor																	
14	*name/title*	*HRS*		*Rate*		*HRS*		*HRS*		*HRS*		*HRS*						
15	Cloud Solution Architect	1200		550	$660,000		$ -		$ -		$ -		$ -	$ -		$	660,000.00	
16	Data Engineer 1	1600		220	$352,000		$ -		$ -		$ -		$ -	$ -		$	352,000.00	
17	Data Engineer 2	1600		220	$352,000		$ -		$ -		$ -		$ -	$ -		$	352,000.00	
18	Data Engineer 3	1400		220	$308,000		$ -		$ -		$ -		$ -	$ -		$	308,000.00	
19	Data Engineer 4	1400		220	$308,000		$ -		$ -		$ -		$ -	$ -		$	308,000.00	
20	ML/AI Engineer 1	1400		250	$350,000		$ -		$ -		$ -		$ -	$ -		$	350,000.00	
21	ML/AI Engineer 2	1400		250	$350,000		$ -		$ -		$ -		$ -	$ -		$	350,000.00	
22	ML/AI Engineer 3	1400		250	$350,000		$ -		$ -		$ -		$ -	$ -		$	350,000.00	
23	Security / Compliance Specialist	1600		320	$512,000		$ -		$ -		$ -		$ -	$ -		$	512,000.00	
24	DevOps/DevSecOps Engineer 1	1600		200	$320,000		$ -		$ -		$ -		$ -	$ -		$	320,000.00	
25	DevOps/DevSecOps Engineer 2	1400		200	$280,000		$ -		$ -		$ -		$ -	$ -		$	280,000.00	
26	Cloud Engineer/Full Stack Engineer 1	1200		200	$240,000		$ -		$ -		$ -		$ -	$ -		$	240,000.00	
27	Cloud Engineer/Full Stack Engineer 2	1200		200	$240,000		$ -		$ -		$ -		$ -	$ -		$	240,000.00	
28	UX Designer	800		250	$200,000		$ -		$ -		$ -		$ -	$ -		$	200,000.00	
29	Business Analyst	1600		200	$320,000		$ -		$ -		$ -		$ -	$ -		$	320,000.00	
30	QA Engineer	1200		200	$240,000		$ -		$ -		$ -		$ -	$ -		$	240,000.00	
31	Functional Consultants/Domain SME	1600		400	$640,000		$ -		$ -		$ -		$ -	$ -		$	640,000.00	
32	Technical Program Manager	1600		450	$720,000		$ -		$ -		$ -		$ -	$ -		$	720,000.00	
33	Total Direct Labor				$6,742,000	0	$ -	0	$ -	0	$ -	0	$ -	$ -		$	6,742,000.00	
34																		
35	II. Direct Labor Fringe Cost	Rate (%):			$0		$ -		$ -		$ -		$ -	$ -		$	-	
36																		
37	II. Subcontractors/Consultants/Contracted Staffing																	
38	*Firm name/title or role/type of services to be provided*	*HRS*		*Rate*		*HRS*		*HRS*		*HRS*		*HRS*						
39					$0		$ -		$ -		$ -		$ -	$ -		$	-	
40					$0		$ -		$ -		$ -		$ -	$ -		$	-	
41					$0		$ -		$ -		$ -		$ -	$ -		$	-	
42					$0		$ -		$ -		$ -		$ -	$ -		$	-	
43					$0	5	$ -		$ -		$ -		$ -	$ -		$	-	
44					$0		$ -		$ -		$ -		$ -	$ -		$	-	
45	Total Subcontractors/Consultants				$0	5	$ -	0	$ -	0	$ -	0	$ -	$ -		$	-	
46																		
47	III. Direct Materials (provide description and costs basis)																	
48	Curriculum Development & Training				$40,000									$0		$	40,000	
49	Infra Cost - GCP				$145,000									$0		$	145,000	
50	Specialized Tooling / Licenses (non-GCP)				$20,000									$0		$	20,000	
51	Documentation Portal & Guides				$10,000									$0		$	10,000	
52	Testing & Benchmark Datasets				$10,000									$0		$	10,000	
53					$0									$0		$	-	
54					$0									$0		$	-	
55	Total Direct Materials				$225,000		$0		$0		$0		$0	$0		$	225,000	
56																		
57	IV. Travel				$351,250									$0		$	351,250	
58																		
59	V. Other Direct Costs *(list type and description such as venue rental costs and cost basis)*																	
60					$0									$ -		$	-	
61					$0									$ -		$	-	
62					$0									$ -		$	-	
63					$0									$ -		$	-	
64					$0									$ -		$	-	
65	Total Other Direct Costs				$0		$ -		$ -		$ -		$ -	$ -		$	-	
66																		
67	VI. General & Administrative Expense/Overhead	Rate (%):			$0		$ -		$0		$0		$0	$ -		$	0	
68																		
69	**Total PROJECT COSTS**				**$7,318,250**		**$0**		**$0**		**$0**		**$0**	**$0**		**$**	**7,318,250**	

LEGEND:

Formula cell

Data entry cell

This section is for post-contract invoicing only. Do not complete with proposal.